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

Факультет социально-экономических и компьютерных науки

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 - N, количество иллюстраций - N, количетсво таблиц - N.

Оглавление

Введение

Этичность компаний уже давно вызывает озабоченность, особенно в отношении их поведения в спорных ситуациях и предоставления услуг, ориентированных на клиента. В последние годы все большее внимание уделяется оценке этичности компаний[1], 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, но эти оценки часто основаны на судебных делах и других официальных отчетах, а не на отзывах клиентов. 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 дать исчерпывающую или точную картину.

Для решения этой проблемы будет реализована система, которая собирала бы и анализировала отзывы потребителей с различных веб-сайтов, чтобы дать более полную и точную оценку этической практики компании. Такая система может быть разработана для автоматического сбора и анализа отзывов потребителей из различных источников, включая социальные сети и сайты отзывов. Затем собранные данные могут быть проанализированы с помощью различных методов, таких как обработка естественного языка и машинное обучение, 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 – деятельность компаний.

Предмет исследования – программные средства для оценки этичности деятельности компаний.

Цель работы – создание системы для оценки этичности компаний.

Исходя из поставленной цели, необходимо:

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

Этап анализа должен:

1. Анализ предметной области
2. 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

Глава 1 Анализ предметной области

1.1. Способы оценки этичности компаний

Компаниям важно оставаться этичными, так как на долгосрочной перспективе это приносит большую прибыль и улучшает показатели бизнеса, чем неэтичный способ ведение бизнеса[2, 1]. На сколько этична компания можно 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[3]

Для пользователей одним из ключевых факторов можно выделить:

- качество пользовательского сервиса[4]
- на сколько навязчивые услуги компании[5]

Кроме того, важно отметить, что 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.

В этой работе для анализа текстов будут использоваться алгоритмы машинного обучения.

1.2. 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[6], TF-IDF[7], Word2Vec[8], ELMO[9], GPT[10] и BERT[11]. 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 своей реализации, но не учитывает порядок слов в предложении. Модель TF-IDF, с другой стороны, представляет текстовые данные, учитывая как частоту слова в документе (TF), так и его важность в общем наборе данных (IDF). Это простые методы анализа текста и не учитывают контекст текста.

Word2Vec, ELMO, GPT и BERT - все это алгоритмы на основе нейронных сетей, которые представляют текстовые данные более сложным способом. Word2Vec представляет слова в виде векторов и может фиксировать значение слов в аналогичных контекстах. ELMO, GPT и BERT основаны на архитектуре transformer, где каждое предложение представлено вектором цифр (эмбеддингом). BERT – лучше остальных алгоритмов понимает текст, так как он может рассматривать слова в контексте всего предложения или текста, когда GPT и ELMO рассматривают только односторонний контекст.

Также для объединения эмбеддинговых пространств будет работать алгоритм подобный CLIP[12], только для трансформации текста в текст.

1.3. Методы

Глава 2 Проектирование системы

2.1. Проектирование базы данных

2.2. Проектирование архитектуры системы

2.2.1. Проектирование серверной части

2.2.2. 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

3.2. Реализация клиентской части

Глава 4 Тестирование системы

Заключение

Библиографический список

1. *Murè P.*, *Spallone M.*, *Mango F.*, *Marzioni S.*, *Bittucci L.* ESG and reputation: The case of sanctioned Italian banks // Corporate Social Responsibility and Environmental Management. — 2021. — Vol. 28, no. 1. — P. 265–277 ; — _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/csr.2047.
2. *Climent F.* Ethical Versus Conventional Banking: A Case Study // Sustainability. — 2018. — July. — Vol. 10, no. 7. — P. 2152 ; — Number: 7 Publisher: Multidisciplinary Digital Publishing Institute.
3. *Harvey B.* Ethical banking: The case of the Co-operative bank // Journal of Business Ethics. — 1995. — Vol. 14, no. 12. — P. 1005–1013.
4. *Brunk K. H.* Exploring origins of ethical company/brand perceptions—A consumer perspective of corporate ethics // Journal of Business Research. — 2010. — Vol. 63, no. 3. — P. 255–262.
5. *Mitchell W. J.*, *Lewis P. V.*, *Reinsch N.* Bank ethics: An exploratory study of ethical behaviors and perceptions in small, local banks // Journal of Business Ethics. — 1992. — Vol. 11, no. 3. — P. 197–205.
6. *Harris Z. S.* Distributional Structure // WORD. — 1954. — Vol. 10, no. 2/3. — P. 146–162.
7. *Jones K. S.* A statistical interpretation of term specificity and its application in retrieval // Journal of documentation. — 1972.
8. *Mikolov T.*, *Sutskever I.*, *Chen K.*, *Corrado G. S.*, *Dean J.* Distributed representations of words and phrases and their compositionality // Advances in neural information processing systems. — 2013. — Vol. 26.
9. *Peters M. E.*, *Neumann M.*, *Iyyer M.*, *Gardner M.*, *Clark C.*, *Lee K.*, *Zettlemoyer L.* Deep contextualized word representations. — 2018.
10. *Radford A.*, *Wu J.*, *Child R.*, *Luan D.*, *Amodei D.*, *Sutskever I.* Language Models are Unsupervised Multitask Learners. — 2019.

11. *Devlin J.*, *Chang M.-W.*, *Lee K.*, *Toutanova K.* Bert: Pre-training of deep bidirectional transformers for language understanding // arXiv preprint arXiv:1810.04805. — 2018.

12. *Radford A.*, *Kim J. W.*, *Hallacy C.*, *Ramesh A.*, *Goh G.*, *Agarwal S.*, *Sastry G.*, *Askell A.*, *Mishkin P.*, *Clark J.*, *Krueger G.*, *Sutskever I.* Learning Transferable Visual Models From Natural Language Supervision // CoRR. — 2021. — Vol. abs/2103.00020.