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 их поведение в спорных ситуациях и предоставление 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]. 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[3].

Для пользователей одними из ключевых факторов можно выделить:

- качество пользовательского сервиса[4];
- насколько 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. Анализ существующих решений

Сейчас данные об этичности компаний можно получить из агрегаторов отзывов и отчётности компаний. Агрегаторы отзывов позволяют собрать информацию о кли-

ентском обслуживании, а отчетность компаний о положении дел в целом. Но сейчас не существует способов, как можно оценить все вместе.

1.3. 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 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 определения 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.

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

1.3.1. BERT

BERT [11] (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 1.1 по механизму описаному в работе [13].

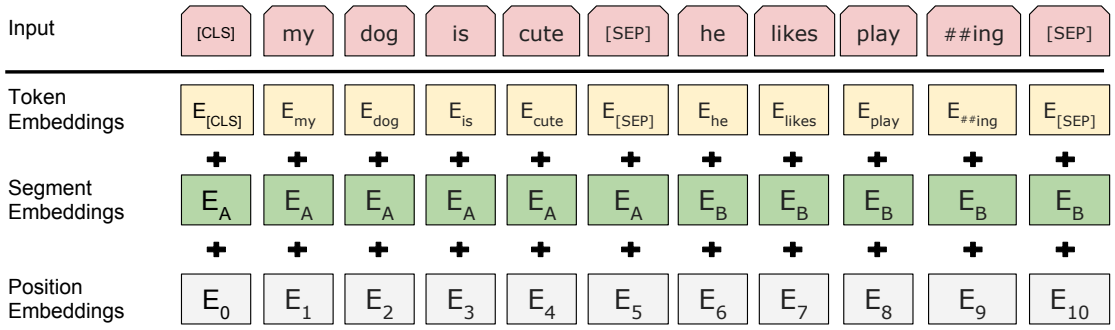

Рисунок 1.1 – Пример ввода текста в модель

При обучении модель выполняет на 2 задания:

1. Предсказание слова в предложении

   Поскольку стандартные языковые модели либо смотрят текст слева направо или справа налево 1.2, как ELMo[9] и GPT[10], они не подходят под некоторые типы заданий. Так как BERT двунаправленный, у каждого слова можно посмотреть его контекст, что позволит предсказать замаскированное слово.

   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,

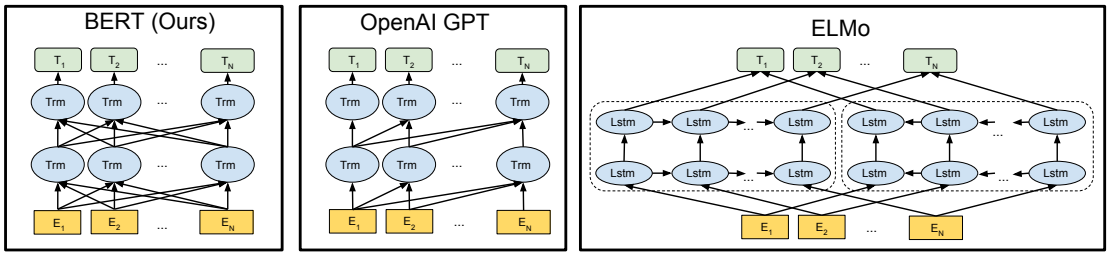

Рисунок 1.2 – Сравнение принципов работы BERT, ELMo, GPT

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 1.3.

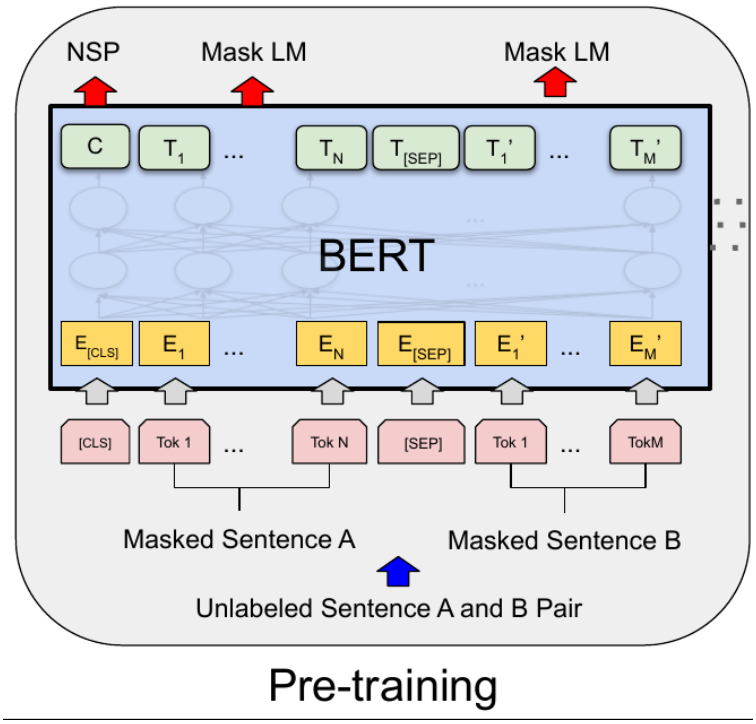

Рисунок 1.3 – Схемам работы BERT

1.3.2. Sentence BERT

Sentense BERT [14] – 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 1.4. Основное преимущество данной

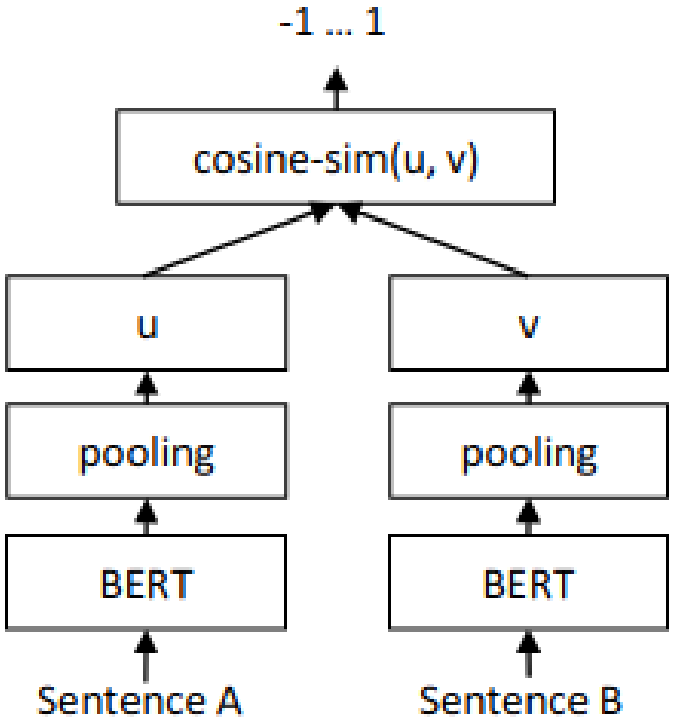

Рисунок 1.4 – Схема работы SBERT

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

1.3.3. CLIP

CLIP (Contrastive Language–Image Pre-training)[12] – это нейронная сеть, обученная на множестве пар (изображение, текст). Его можно проинструктировать на естественном языке, чтобы он предсказал наиболее релевантный фрагмент текста, учитывая изображение, без прямой оптимизации для задачи. Эта модель состоит из двух разных моделей. Одна для кодирования текста в эмбеддинг – трансформер [13], а для кодирования изображения используется vision transformer [15]. В данной работе будет использована модификация этого метода для сопоставления текстов из разных сфер между собой.

Метод обучения данной модели авторы отнесли к "natural language supervision"(обучение естественным языком). На каждой итерации обучения берется набор пар изображение-текст. Затем они трансформируются в эмбеддинги и к каждому тексту модель пытается подобрать текст, и наоборот. Данный способ позволяет соединить пространства двух различных источников информации.

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

3.1.5. Дообучение модели

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.

13. *Vaswani A.*, *Shazeer N.*, *Parmar N.*, *Uszkoreit J.*, *Jones L.*, *Gomez A. N.*, *Kaiser Ł.*, *Polosukhin I.* Attention is All you Need // Advances in Neural Information Processing Systems. Vol. 30. — Curran Associates, Inc., 2017.

14. *Reimers N.*, *Gurevych I.* Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. — Association for Computational Linguistics, 11/2019.

15. *Dosovitskiy A.*, *Beyer L.*, *Kolesnikov A.*, *Weissenborn D.*, *Zhai X.*, *Unterthiner T.*, *Dehghani M.*, *Minderer M.*, *Heigold G.*, *Gelly S.*, [et al.]. An image is worth 16x16 words: Transformers for image recognition at scale // arXiv preprint arXiv:2010.11929. — 2020.