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

При работе с различными компаниями возникают проблемы их надежности, то как они ведут себя в спорных ситуациях, есть ли сервисы направленные на взаимодействие с клиентами. Обращаясь к различным работам [1, 2, 3], можно увидеть, что оценка этичности компаний, в данных работах на примере банковской сферы и ESG фактора, в настоящее время очень актуальна. В данное время существуют сервисы, которые могут оценить этичность компании на основании судебных дел и общей оценки состояния компаний, но не на отзывах о них. Из-за этого людям при выборе компаний приходится смотреть отзывы с различных сайтов об их услугах и самому анализировать насколько этична компания. Для решения этой проблемы будет реализована система, которая будет собирать отзывы с различных сайтов и анализировать их.

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. BERT

BERT [4] (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. В отличие от TF-IDF [5], мешка слов[6] и word2vec [7] последовательности слов и предложений кодируются вектором (эмбеддингом) фиксированной длины. Она позволяет эффективно анализировать текст с пониманием контекста, что поможет при анализе отзывов.

На вход модели подается предложение или пара предложений. Затем разделяется на отдельные слова (токены). Потом в начало последовательности вставляется специальный токен `[CLS]`, обозначающий начало предложения или начало последовательности предложений. Пары предложений группируются в одну последовательность и разделяются с помощью специального токена `[SEP]`. Потом все токены превращаются в эмбеддинги 1.1 по механизму внимания [8].

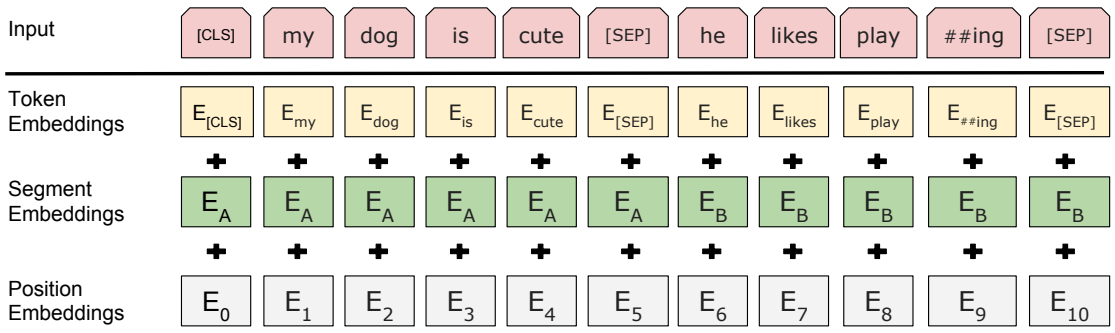

Рисунок 1.1 – Пример ввода текста в модель

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 работают с контекстом хуже, чем данная модель. Так как BERT двунаправленный, у каждого слова можно посмотреть его контекст, что позволит предсказать замаскированное слово.

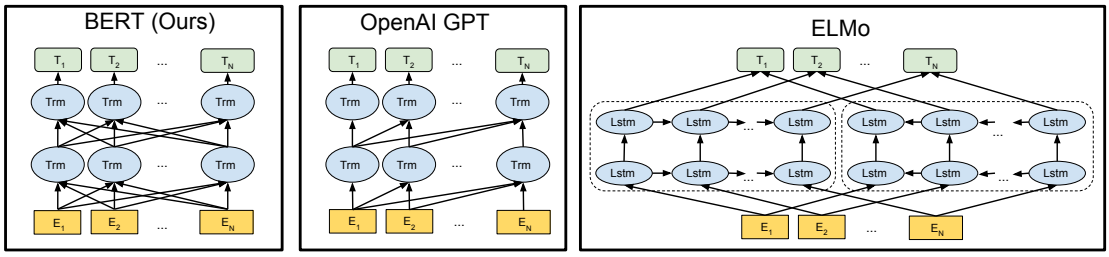

Рисунок 1.2 – Сравнение принципов работы BERT, ELMo, GPT

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 1.3.

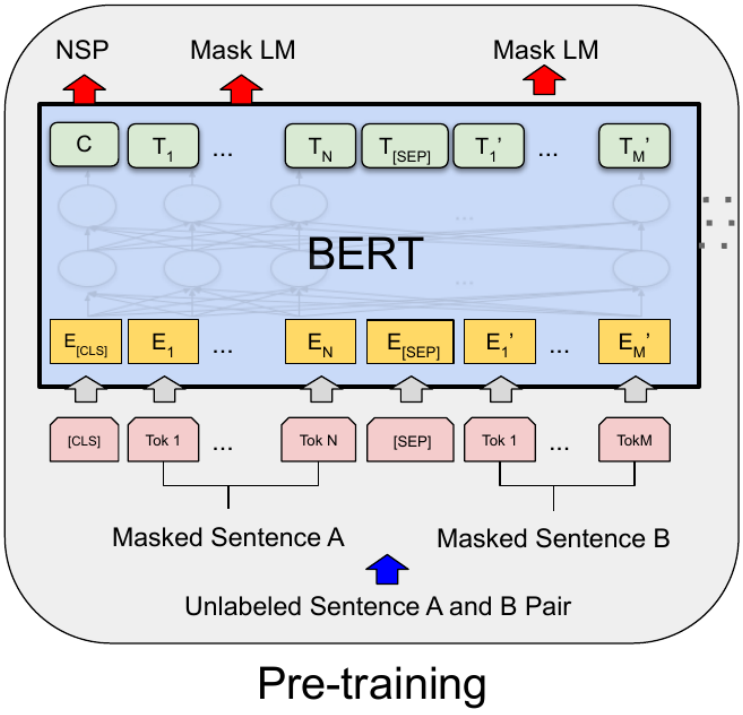

Рисунок 1.3 – Схемам работы BERT

1.2. Sentense BERT

Sentense BERT [11] – это модификация предобученных моделей BERT, которая использует модель BERT и подает на вход 2 предложения, затем усреднят их выходы, а после с помощью функции ошибки выдаёт результат. Схема работы модели 1.4. Основ-

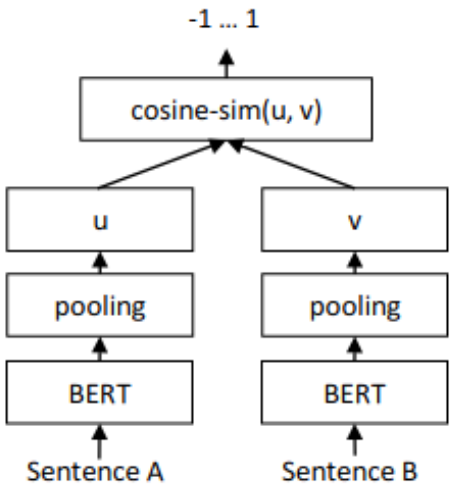

Рисунок 1.4 – Схема работы SBERT

ное 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 рассчитает эмбеддинг каждого предложения отдельно и потом их сравнит, и это займёт примерно 5 секунд.

1.3. 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 [8], а для кодирования изображения используется vision transformer [13]. 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

3.2. 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. — Т. 28, № 1. — С. 265—277 ; — _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/csr.2047.

2. *Miralles-Quirós M. M.*, *Miralles-Quirós J. L.*, *Redondo Hernández J.* ESG Performance and Shareholder Value Creation in the Banking Industry: International Differences // Sustainability. — 2019. — Янв. — Т. 11, № 5. — С. 1404 ; — Number: 5 Publisher: Multidisciplinary Digital Publishing Institute.

3. *Climent F.* Ethical Versus Conventional Banking: A Case Study // Sustainability. — 2018. — Июль. — Т. 10, № 7. — С. 2152 ; — Number: 7 Publisher: Multidisciplinary Digital Publishing Institute.

4. *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.

5. *Jones K. S.* A statistical interpretation of term specificity and its application in retrieval // Journal of documentation. — 1972.

6. *Harris Z. S.* Distributional Structure // WORD. — 1954. — Т. 10, № 2/3. — С. 146—162.

7. *Mikolov T.*, *Chen K.*, *Corrado G.*, *Dean J.* Efficient estimation of word representations in vector space // arXiv preprint arXiv:1301.3781. — 2013.

8. *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. Т. 30. — Curran Associates, Inc., 2017.

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. *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.

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. — T. abs/2103.00020.

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