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

Рецензент

I.I. Ivanov

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 4

Глава 1 Анализ предметной области 6

1.1 Постановка задачи 6

1.2 BERT 6

1.3 Sentense BERT 7

Глава 2 Проектирование системы 9

2.1 Проектирование базы данных 9

2.2 Проектирование архитектуры системы 9

2.2.1 Проектирование серверной части 10

2.2.2 Проектирование клиентской части 10

Глава 3 Реализация системы 11

3.1 Реализация серверной части 11

3.1.1 Реализация API 11

3.1.2 Реализация парсера banki.ru 11

3.1.3 Реализация парсера sravni.ru 11

3.1.4 Реализация модуля обработки текста 11

3.2 Реализация клиентской части 11

Глава 4 Тестирование системы 12

Заключение **13**

Библиографический список **14**

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.

В настоящее время существуют сервисы, которые могут оценить этичность компании на основании судебных дел, но не на отзывах о компании.

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. Постановка задачи

В данный момент при выборе компаний приходится смотреть отзывы на различных сайтов и самому анализировать на сколько этична компания.

1.2. BERT

BERT [1] (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 [2].

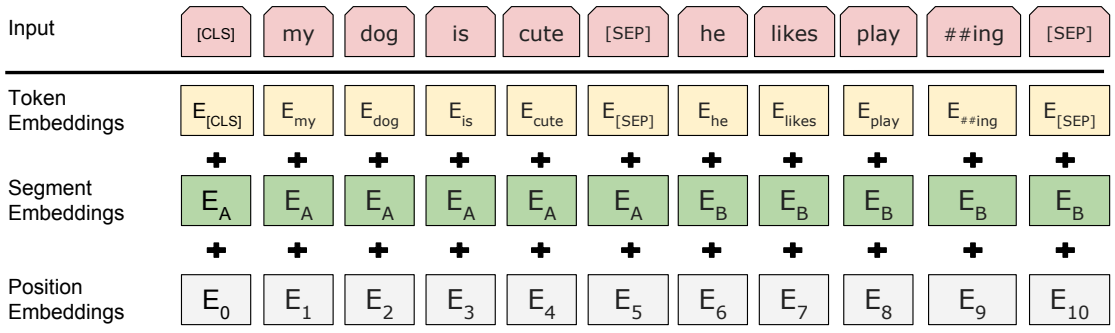


Рисунок 1.1 – Пример ввода текста в модель

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[3] и GPT[4], они не подходят под некоторые типы

заданий. Так как BERT двунаправленный, у каждого слова можно посмотреть его контекст, что позволит предсказать замаскированное слово.

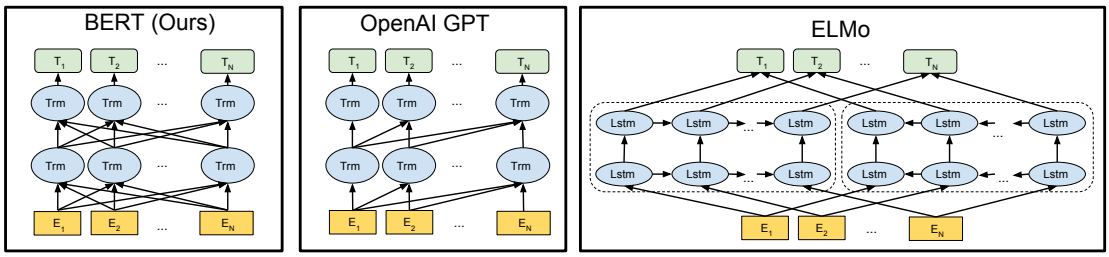


Рисунок 1.2 – Сравнение принципов работы BERT, ELMo, GPT

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.

1.3. Sentense BERT

Sentense BERT [5] – 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 1.4. 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 рассчитает эмбеддинг каждого предложения отдельно и потом их сравнит, и это займёт примерно 5 секунд.

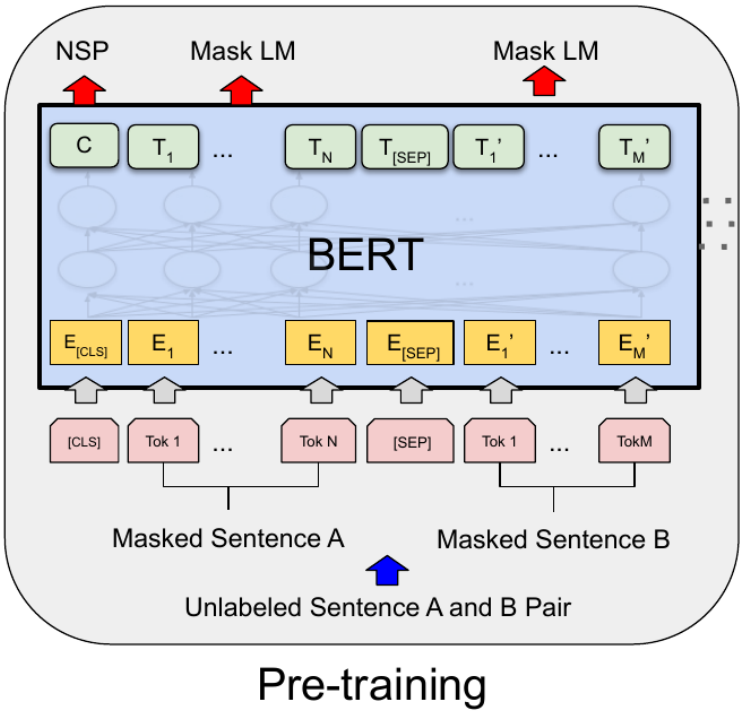


Рисунок 1.3 – Схемам работы BERT

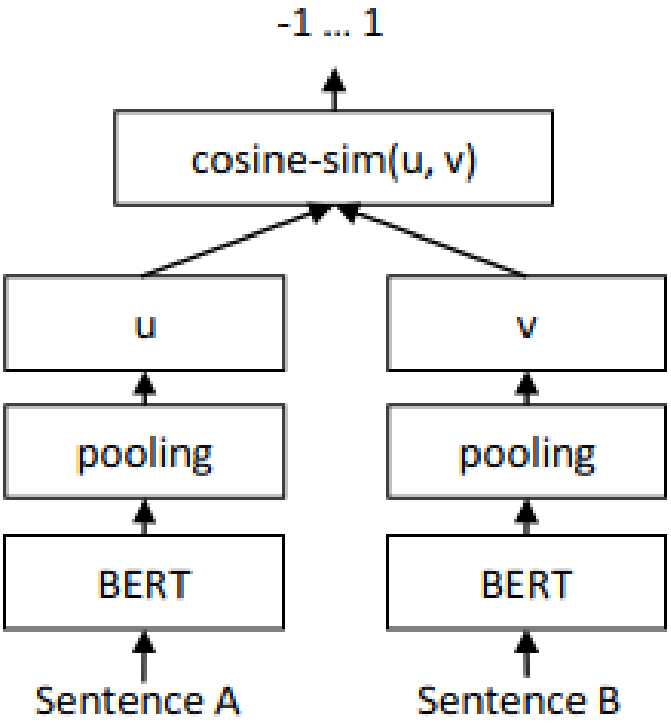
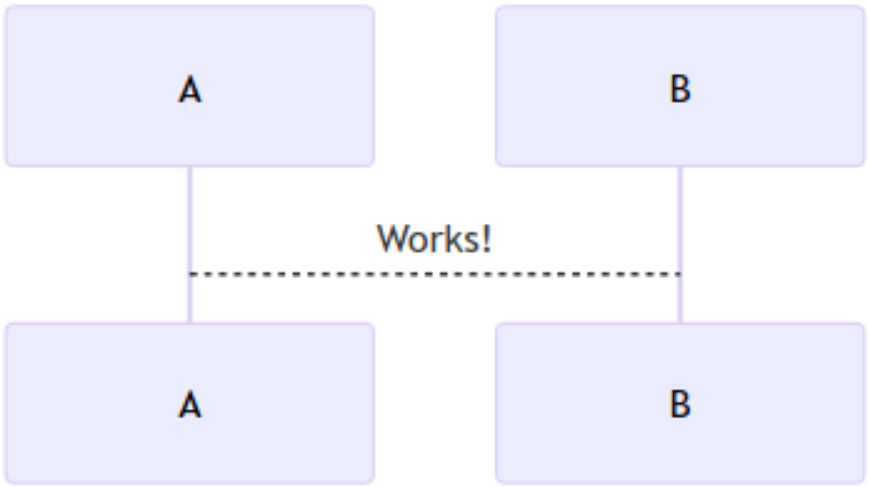


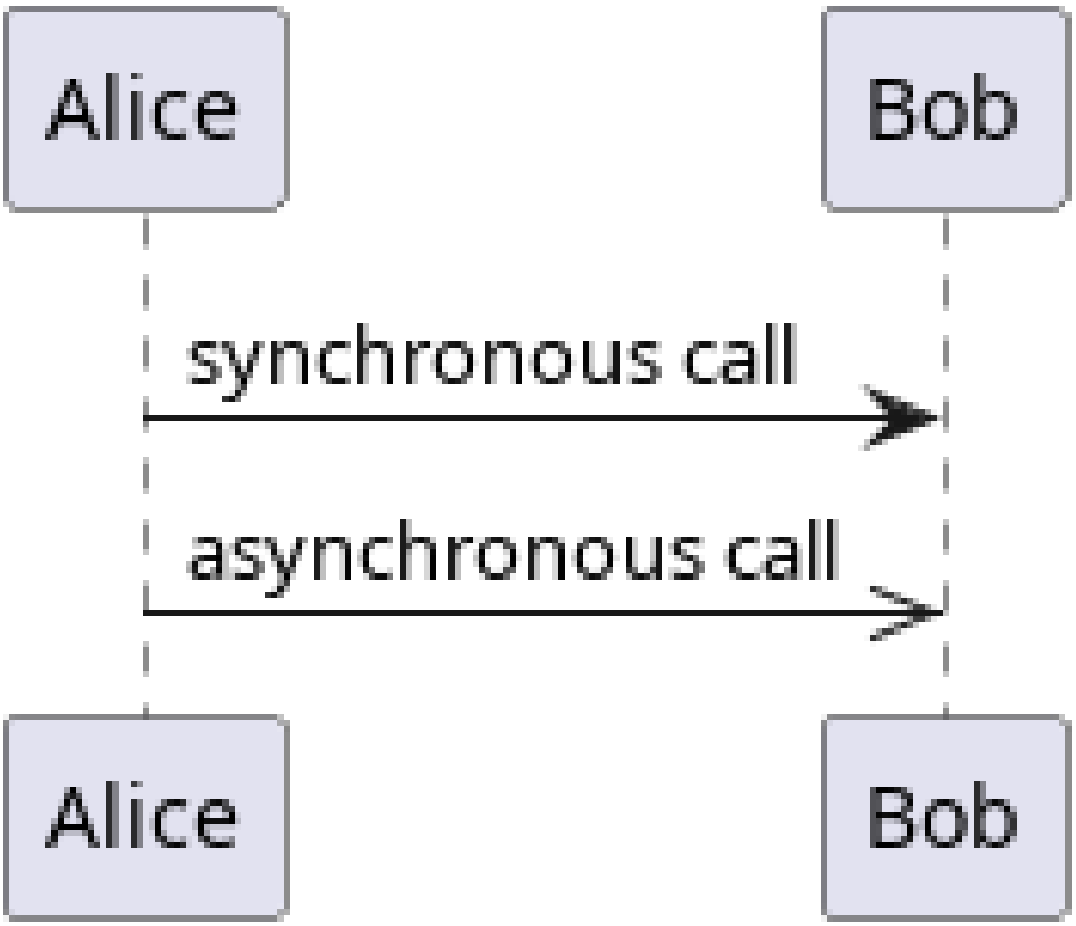
Рисунок 1.4 – Схема работы SBERT

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. *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.
2. *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. T. 30. — Curran Associates, Inc., 2017.
3. *Peters M. E.*, *Neumann M.*, *Iyyer M.*, *Gardner M.*, *Clark C.*, *Lee K.*, *Zettlemoyer L.* Deep contextualized word representations. — 2018.
4. *Radford A.*, *Wu J.*, *Child R.*, *Luan D.*, *Amodei D.*, *Sutskever I.* Language Models are Unsupervised Multitask Learners. — 2019.
5. *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.
6. *Кафедра ИТБ НИУ ВШЭ-Пермь*. Курсовые работы и ВКР. — 2020. — URL: `https://www.hse.ru/data/2020/11/26/1350963672/%D0%9F%D1%80%D0%B0%D0%B2%D0%B8%D0%BB%D0%B0%20%D0%92%D0%9A%D0%A0%20%D0%9F%D0%98%20(11.2020).pdf` (дата обр. 13.11.2022).
7. *Kuratov Y.*, *Arkhipov M.* Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. — 2019.