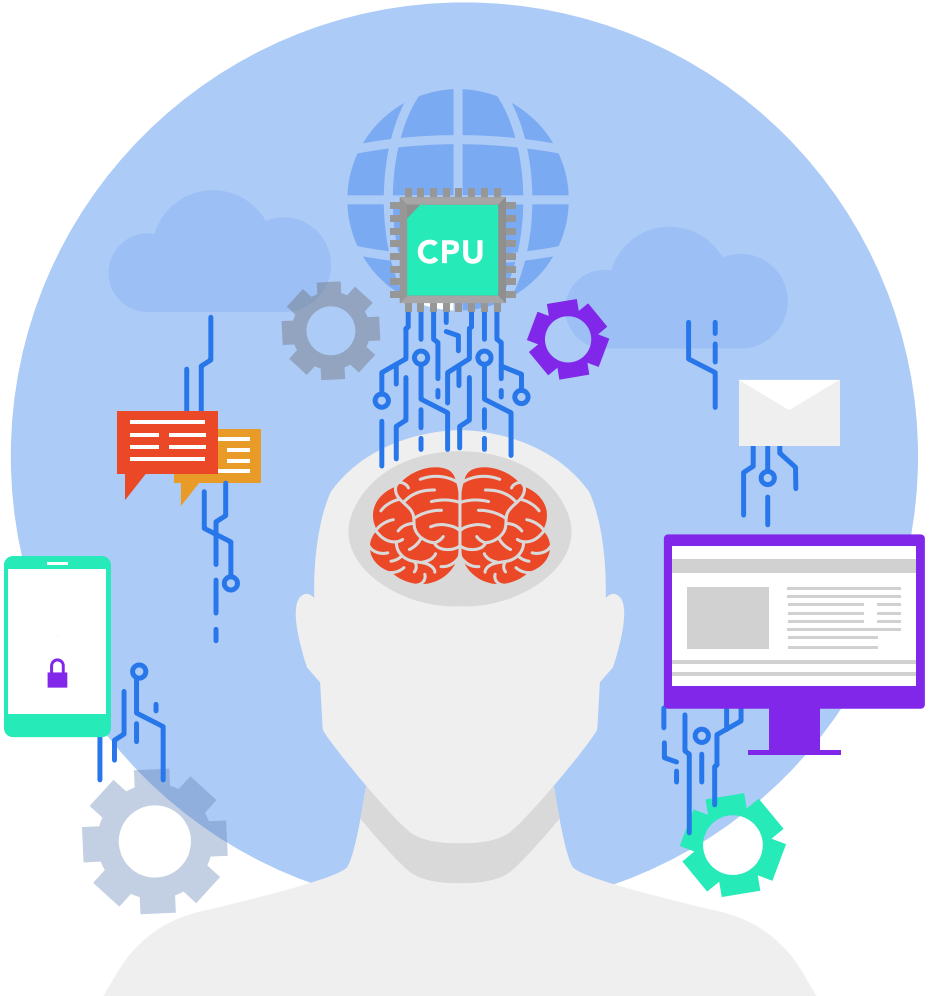
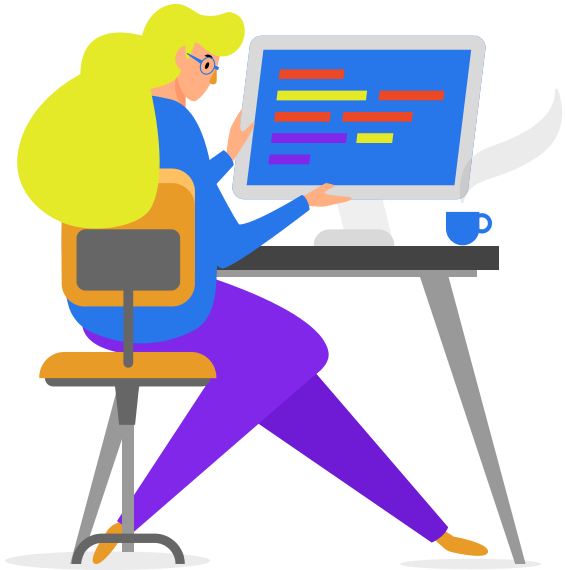


Cup IT 2023
Трек: Data science

TeamRoulette
команда №411

Намеченный план



01 Направление
Имеющийся опыт
Роли
Чекпоинты

02 Анализ
Данные
Гипотезы
Схожие кейсы

03 Выбор приоритетов
Дальнейший план
Основные аспекты

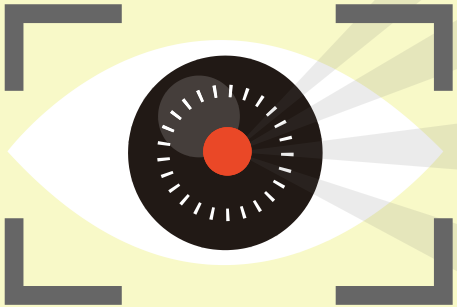
04 Разработка решения
Проверка гипотез
Обучение моделей

05 Анализ результатов
Сравнение
Отчёт
Доработка

06 Оформление решения
Выводы
Презентация

Направление

Основные выдвинутые гипотезы



Содержание

Ссылки, числа и объем комментария могут значительно влиять на ранкинг

Современный подход

Попробовать современные NLP модели, такие как трансформеры

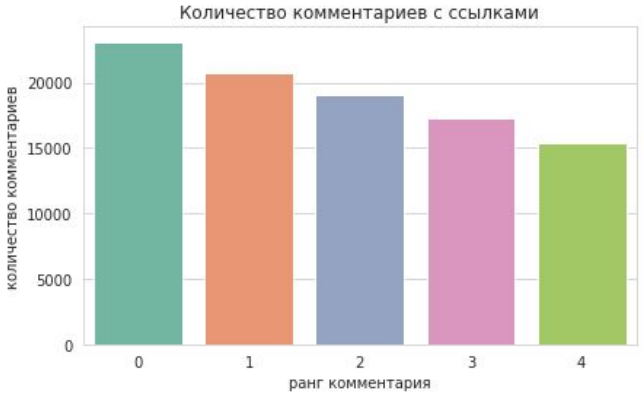
Работа в паре

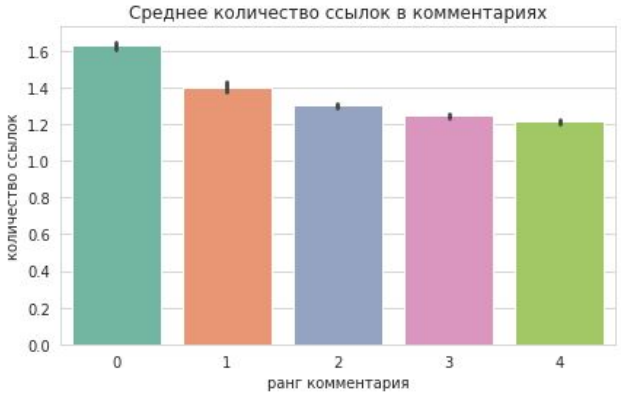
Отправлять в модель текст и комментарий, использовать sentence-level classification

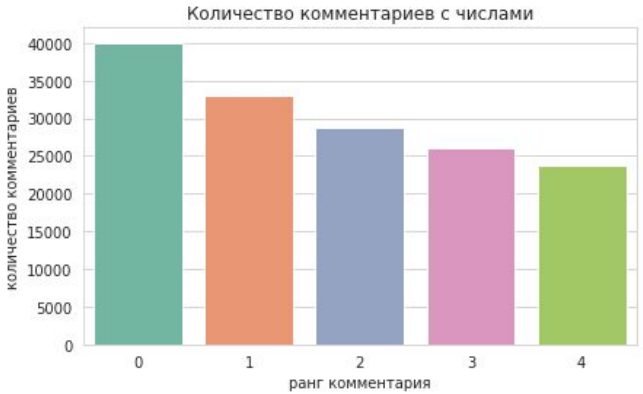
Ансамбль моделей

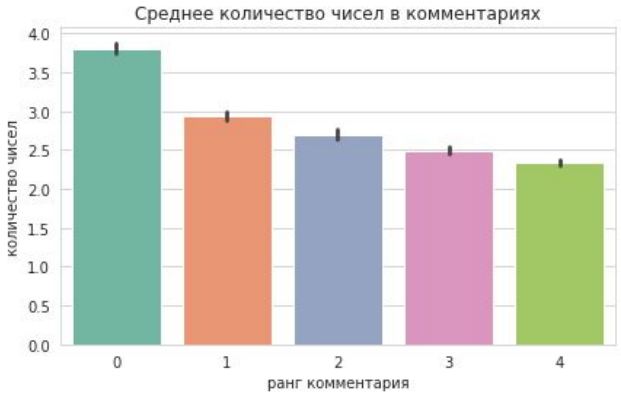
Ансамбль моделей может существенно улучшить качество ранжирования

Влияние ссылок и чисел









На ранг комментария влияет не только наличие ссылок или чисел в нем, но также и их количество.

В популярных комментариях чаще встречаются ссылки и числа и, в среднем, количество символов больше

Дизайн модели

Проработка датасета

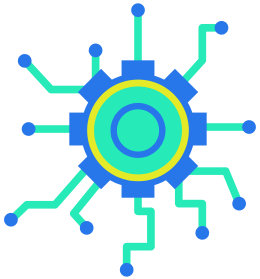
- Датасет возможно настроить на получение любого количества данных, как то любое количество постов, комментариев к ним, так и любое количество рангов

Vs

Тестирование моделей

- В ходе изучения существующих подходов и экспериментов над предобученными BERT, DistelBERT, XLM моделями было выявлено, что для данной работы по занимаемой памяти, скорости и качеству предсказания наиболее подходит предобученная DistelBert-based модель

Дообучение DistilBert

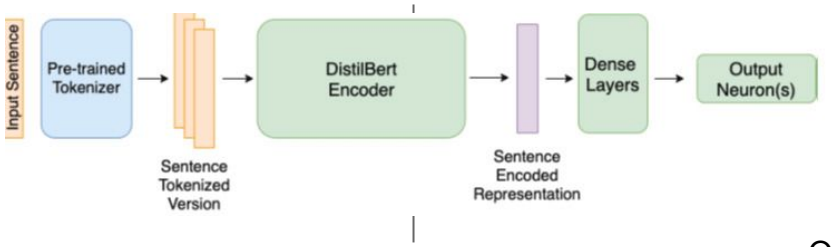


Архитектура

Механизм внимания, на котором основаны трансформеры, позволяет анализировать последовательность элементов, а не отдельные части

Обработка данных

К виду [CLS] TEXT [SEP] COMMENT [SEP]
Используем предобученный токенизатор



Предсказание

Ранжирование по значению выходного нейрона

MSELoss и NDCG метрика

Оцениваем, насколько близки предсказанные значение к лейблам и как соблюдается порядок

Результаты

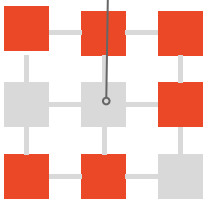
01

Проведены исследования

В области обработки естественных языков, в частности, ранжирования комментариев

github

02

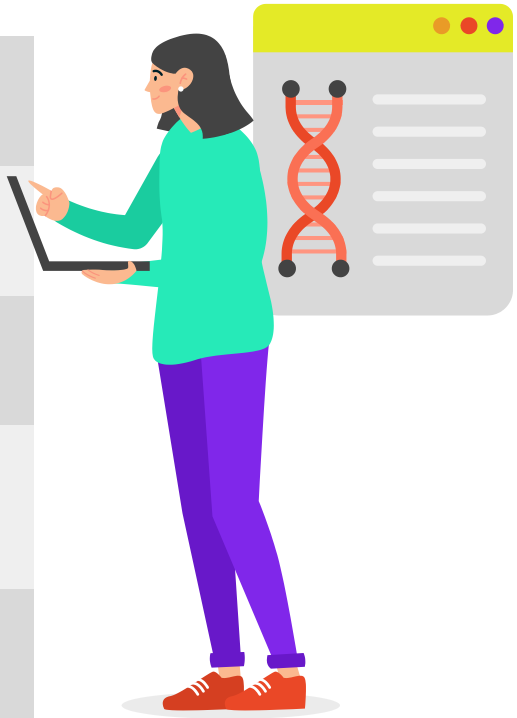


Получена модель

Осуществляющая ранжирование с метрикой NDCG 0.97 на валидационной выборке

Рекомендации пользователям

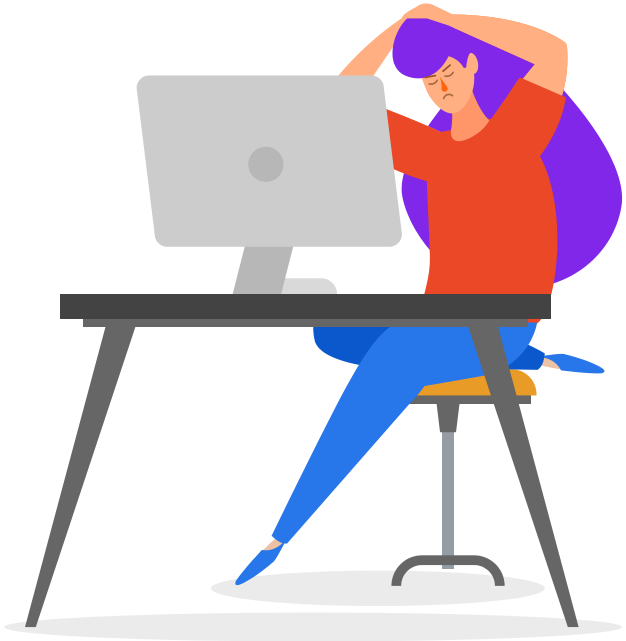
01	Длина комментария	Развернутые комментарии более популярны
02	Ссылки	Комментарии с несколькими ссылками ранжируются выше
03	Числа	Аналогично ссылкам, числа повышают релевантность
04	Совпадения	Использование в комментарии слов из поста повышает его ранг
05	Email	Наличие контактной информации также повышает ранг



Методы взаимодействия с комментаторами

Предложение пользователям различных рекомендаций на этапе написания комментария, чтобы повысить его популярность
Например, раскрыть идею комментария более подробно или добавить ссылку на дополнительный материал по теме

Спрашивать у некоторых пользователей, которые заходят вглубь списка комментариев, с помощью диалогового окна, согласны ли они с тем, что верхний ответ самый полезный/интересный



Наша команда

Дерунец Роман

НГУ, ИИР, 3 курс
Роль:
Обучение моделей, NLP, оформление

Сергей Присяжный

СГТУ, выпускник
Роль:
ML, DL дизайн, обучение моделей

Кутняк Юлия

ФУ, ИТиАБД, 3 курс
Роль:
Обработка данных, анализ