

Кейс:

Центральный Банк Российской Федерации

Анализатор текстовых пресс-релизов

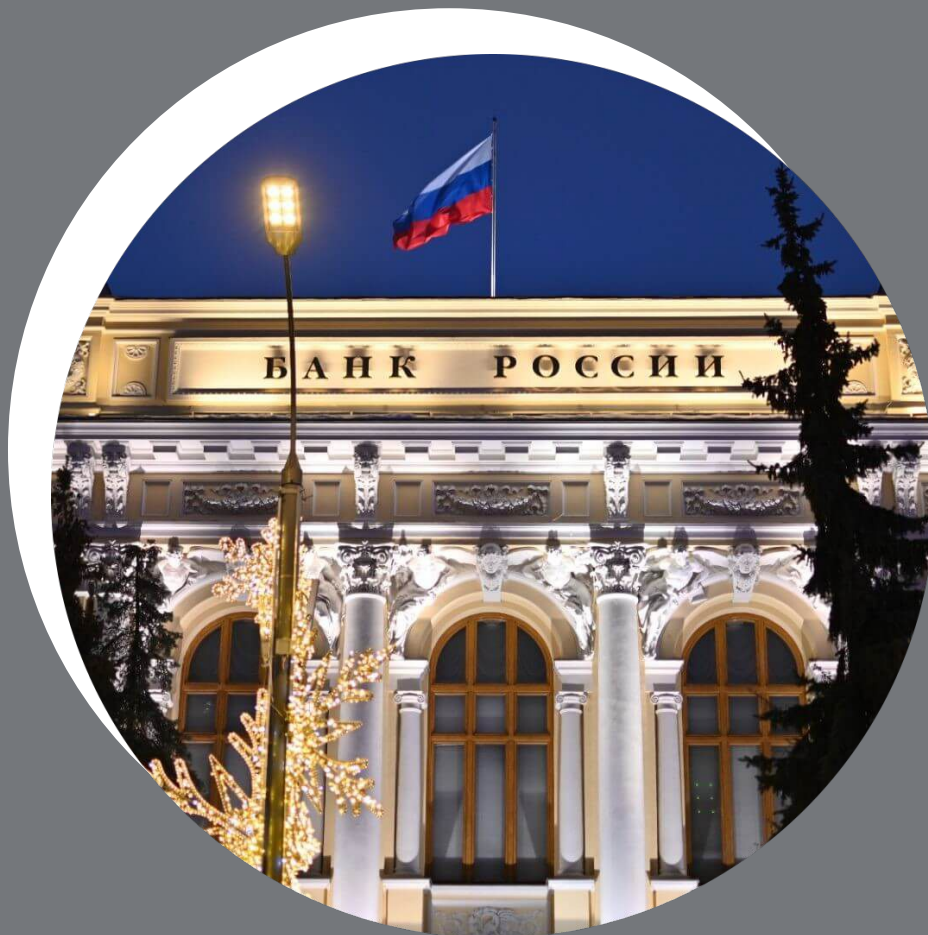
Стек технологий:

Python, PyTorch, TensorFlow, Pandas, NumPy,
Matplotlib, BERT, Natasha, Sklearn



СОЮЗ

@domnenko_a_n
0001@list.ru



РЕШЕНИЕ

BERT

Fine tuning

точная настройка
предобученной модели,
дообучение на своих
данных



+

Killer

Features

Text + NER

count	average	unique	name	data	loc
-------	---------	--------	------	------	-----

Concatenate

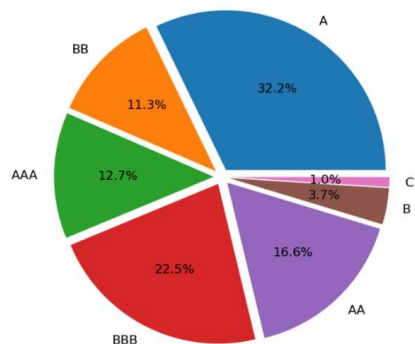


F1 Score + 2-3%

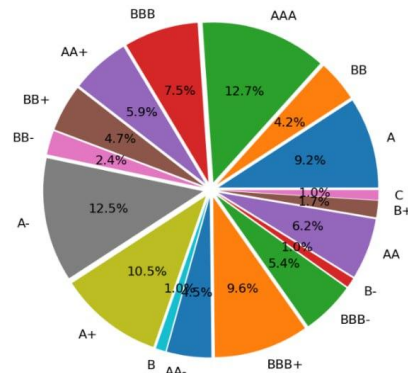
F1 Score - 0,6579 / 0.4778

АНАЛИЗ И ОБРАБОТКА ДАННЫХ

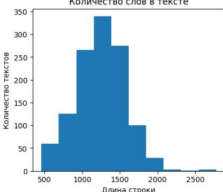
Распределение категорий



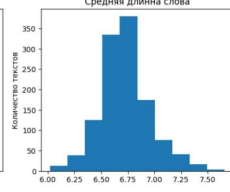
Распределение рейтингов



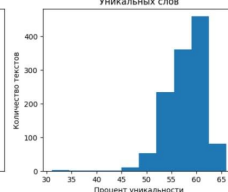
Количество слов в тексте



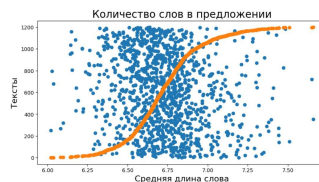
Средняя длина слова



Уникальных слов




Количество слов в предложении



Удаляем:

- `\n\t`
- `<теги>`
- ссылки `http://`
- спецсимволы
- двойные пробелы
- NER
 - ('АО «Эксперт РА', 10416),
 - ('АКРА', 3575),
 - ('Компания', 2997),
 - ('Группа', 1807)

МАСШТАБИРУЕМОСТЬ

Технические	Бизнес
<p data-bbox="85 369 369 405">Пути улучшения</p> <ul data-bbox="112 456 919 918" style="list-style-type: none"><li data-bbox="112 456 919 798">● Работа с NER (Deerpravllov) :<ul data-bbox="208 500 919 798" style="list-style-type: none"><li data-bbox="208 500 919 532">○ увеличение количества сущностей,<li data-bbox="208 532 919 565">○ разбиение сущностей на классы<li data-bbox="208 565 919 598">○ интерпретируемость фич<li data-bbox="208 598 919 685">○ интерпретируемость признаков с помощью шейп<li data-bbox="208 685 919 718">○ первые слова по распределению ципфа<li data-bbox="208 718 919 798">○ для сущностей находим зависимые слова в предложениях<li data-bbox="112 798 919 918">● Работа с текстом<ul data-bbox="208 841 919 918" style="list-style-type: none"><li data-bbox="208 841 919 918">○ подсчет количества иностранных слов и др.	<p data-bbox="966 369 1338 405">Области применения</p> <ul data-bbox="993 456 1725 721" style="list-style-type: none"><li data-bbox="993 456 1725 721">● Работа с внешними факторами:<ul data-bbox="1089 500 1725 721" style="list-style-type: none"><li data-bbox="1089 500 1725 532">○ оценка тональности комментариев<li data-bbox="1089 532 1725 565">○ определение настроения инвесторов<li data-bbox="1089 565 1725 685">○ предсказание тренда роста рынка на основе переписки инвесторов в общедоступных чатах<li data-bbox="1089 685 1725 721">○ предсказательная аналитика 

ООО ТК Нафаттранс Плюс |
Российская Федерация |
Новосибирск |
Агентство АКРА | ПАО |
ВымпелКом |
ООО РКС Холдинг |
Агентство | Петрозаводск |
Благовещенск | Фонд
развитие территории |
АО ГК ДИНАМИКА |
АО ГК Азот | Москва |
Китай | Тайвань |

НАША КОМАНДА



Хуторной Борис

Data scientist
front-end



Чесников Леонид

Data scientist



Ярулин Дамир

Data scientist
front-end



Василевский Иван

Data scientist



Домненко Алексей

Data scientist
Full-stack

