

# Домашнее задание по теме «Трансформер, BERT»

## Формулировка задания

Использовать предварительно обученную модель BERT для классификации текста.

*Обратить внимание, что большинство этапов домашнего задания представлены в лекции по Трансформерам и BERT.*

*Пункт, который отмечен как “Дополнительно”, можно пропустить при выполнении.*

## План работы

- 1) Загрузить набор данных настоящих и поддельных новостей с kaggle: <https://www.kaggle.com/datasets/nopdev/real-and-fake-news-dataset>. Датасет содержит метку правдивости новости, заголовок и текст новости.
- 2) Провести очистку данных: убрать пунктуацию и стоп-слова. Привести текст к нижнему регистру. При необходимости, заменить пробелы символом “\_”.
- 3) Разделить данные на тренировочную, тестовую выборки.
- 4) Использовать библиотеку transformers и модель Hugging Face для загрузки предварительно обученной модели BERT и токенизатора.
- 5) Подготовить данные: использовать токенизатор BERT для преобразования текстовых данных в формат, который можно подать на вход модели BERT.
- 6) Создать классификатор на основе BERT: это может быть модель BERT с одним линейным слоем для классификации на вершине.
- 7) Обучить классификатор на тренировочных данных новостей и оценить его производительность на данных для тестирования.

8) Дополнительно. Изменить гиперпараметры модели и переобучить классификатор. Сделать вывод о том, как изменилось качество классификации.

## Перечень инструментов, необходимых для реализации деятельности

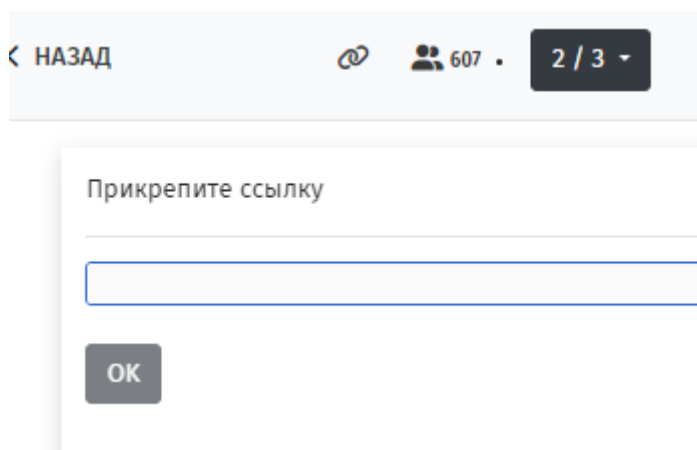
Google Colab <https://colab.research.google.com/>

Библиотека **transformers**

Библиотеки **nlk**, **sklearn**, **torch**, **pandas**, **numpy**

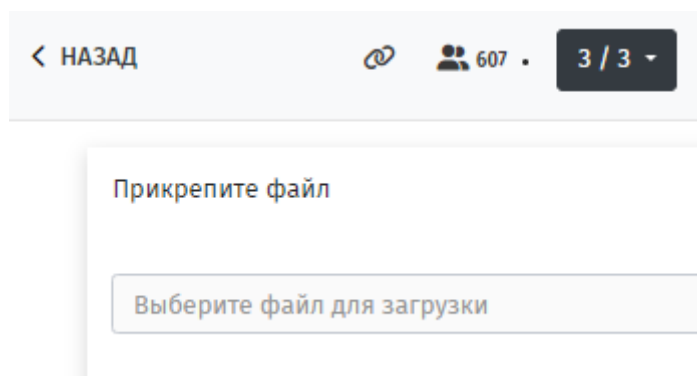
## Форма загрузки

- В поле ссылки (2 страница задания) загрузить ссылку на ноутбук google colab или github репозиторий.



The screenshot shows the Google Colab interface. At the top, there is a navigation bar with a back arrow, the text 'НАЗАД', a share icon, a user icon with '607', and a tab indicator '2 / 3'. Below this, a modal dialog box titled 'Прикрепите ссылку' (Attach link) is displayed. It contains a text input field and an 'ОК' button.

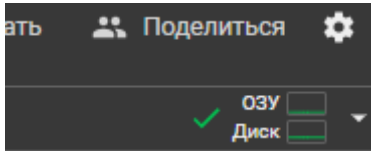
- В поле файла (3 страница задания) загрузить ноутбук с решением (файл с расширением .ipynb).



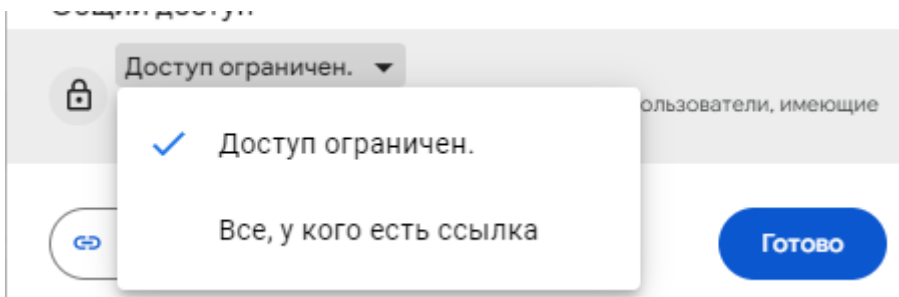
The screenshot shows the Google Colab interface. At the top, there is a navigation bar with a back arrow, the text 'НАЗАД', a share icon, a user icon with '607', and a tab indicator '3 / 3'. Below this, a modal dialog box titled 'Прикрепите файл' (Attach file) is displayed. It contains a text input field with the placeholder text 'Выберите файл для загрузки' (Select file for upload).

## Инструкция по получению ссылки на ноутбук google colab

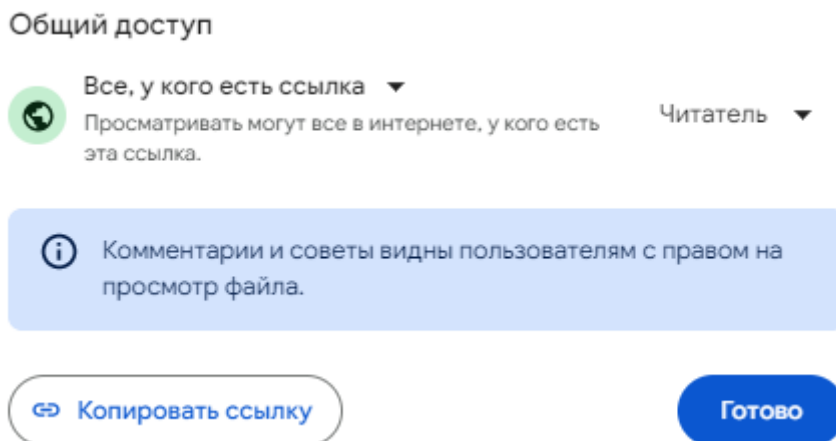
1. Нажмите **“Поделиться”** в правом верхнем углу экрана, рядом с лого вашего google аккаунта



2. В поле **“Общий доступ”** вместо **“Доступ ограничен”** выберите **“Все у кого есть ссылка”**

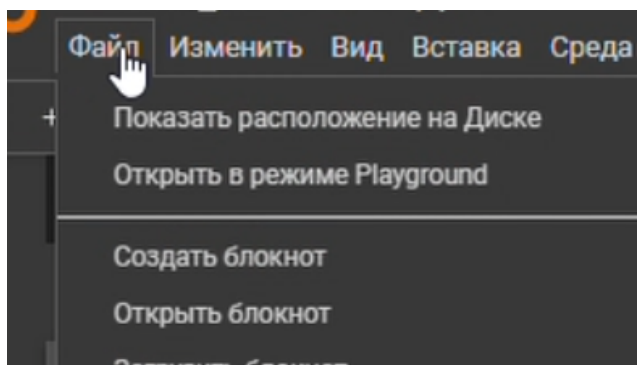


3. Нажмите **“Копировать ссылку”** и вставьте ее в поле ссылки

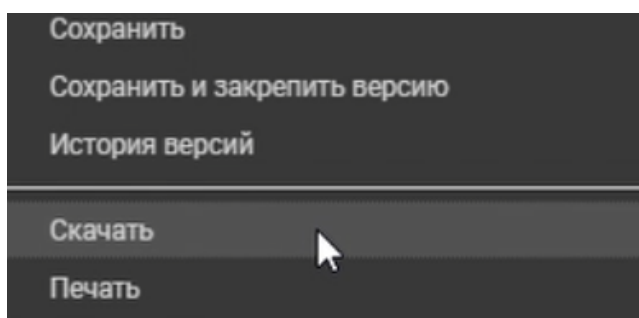


## Инструкция по скачиванию файла с google colab

В меню **“Файл”**



Выбрать пункт **“Скачать”**



Выбрать пункт **“IPYNB”**

