**Информация о датасетах**

**Что известно о датасете Satellites:**

Satellites – папка с датасетами.

В ней есть 2 папки: train и test.

train – Данные используемые нами в обучении нейросети.

Эссе хранятся в .json файлах.

Результат наличия или отсутствия фактической ошибки во всех эссе содержится в файле train\_standart.json.

Test – Данные для теста скачанные с платформы.

test\_task.json – файл, который нужно будет заполнить результатом работы нейросети.

Эссе для теста хранятся в .json файлах.

**Наш анализ датасета Satellites:**

С помощью специального колаба был проведен анализ датасета Satellites.

Результаты анализа были записаны в файл:

dataset\_analysis\_result.json

Каждому эссе соответствует верный ответ true или false, говорящий о наличии или отсутствии фактических ошибок в эссе. (Чему соответствует true, a чему false – не известно).

В датасете есть дисбаланс по кол. эссе с фактическими ошибками (True) и без них (False).

Возможное решение дисбаланса:

Можно сделать аугментацию текста, а можно урезать false.

Однако урезание false будет плохим, т.к. останется очень мало данных.