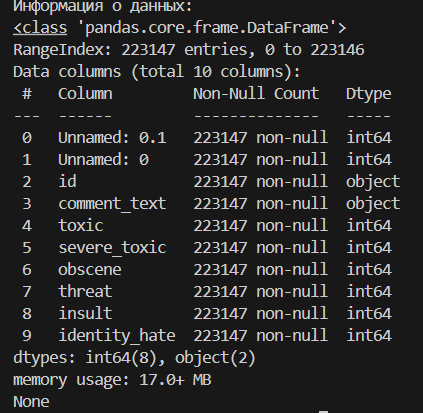
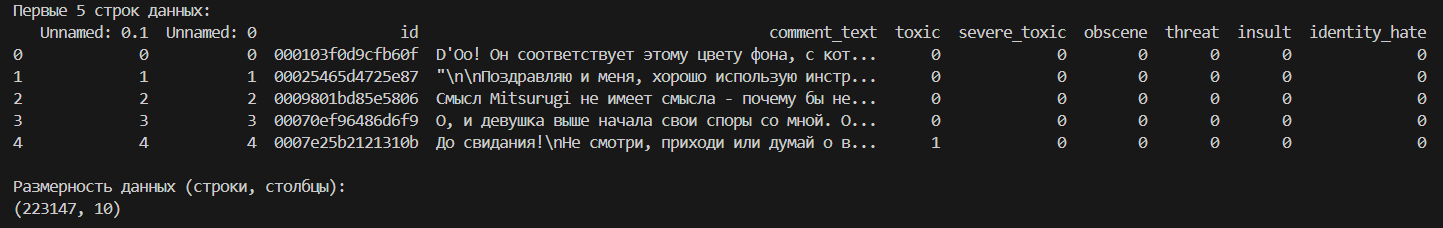
1. Сбор данных

|  |  |
| --- | --- |
| Источник | Ссылка |
| jigsaw-toxic-comment-classification-challenge | https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data |
| jigsaw-unintended-bias-in-toxicity-classification | https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/data |
| jigsaw-multilingual-toxic-comment-classification | https://www.kaggle.com/c/jigsaw-multilingual-toxic-comment-classification/data |
|  | <https://www.kaggle.com/datasets/ludovick/jigsawtanslatedgoogle>  ?select=jigsaw-unintended-bias-train\_ru\_clean.csv |
|  | https://www.kaggle.com/datasets/miklgr500/jigsaw-train-multilingual-coments-google-api |

1. Очистка

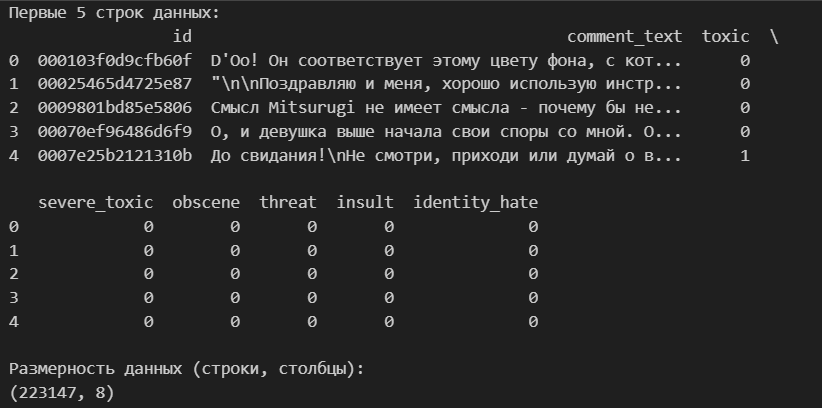
C:\Users\bokar\Documents\GitHub\AI-Stream\detoxify\jigsaw\_data\jigsaw-multilingual-toxic-comment-classification\ jigsaw-toxic-comment-train-google-ru-cleaned.csv

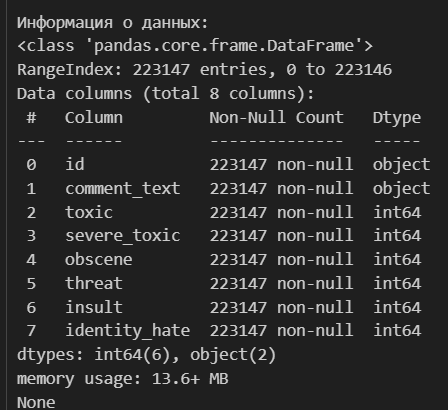




Столбцы Unnamed: 0.1 и Unnamed: 0 выглядят как технические (возможно, индексы, оставшиеся после сохранения данных)

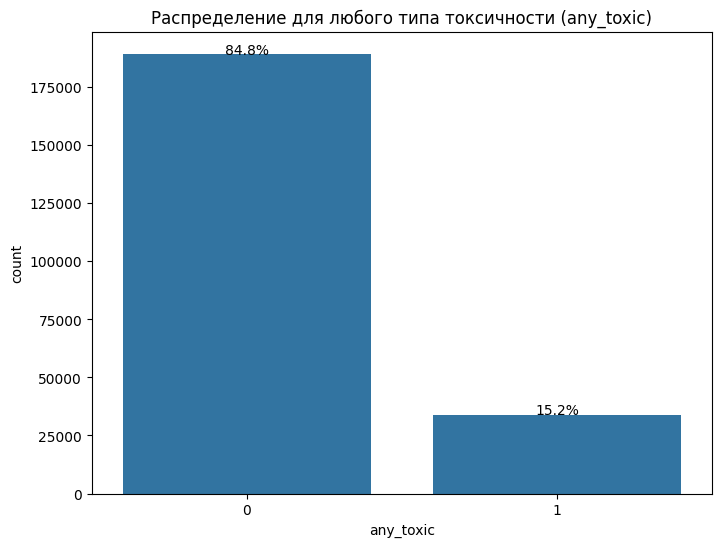
df.drop(columns=['Unnamed: 0.1', 'Unnamed: 0'], inplace=True)

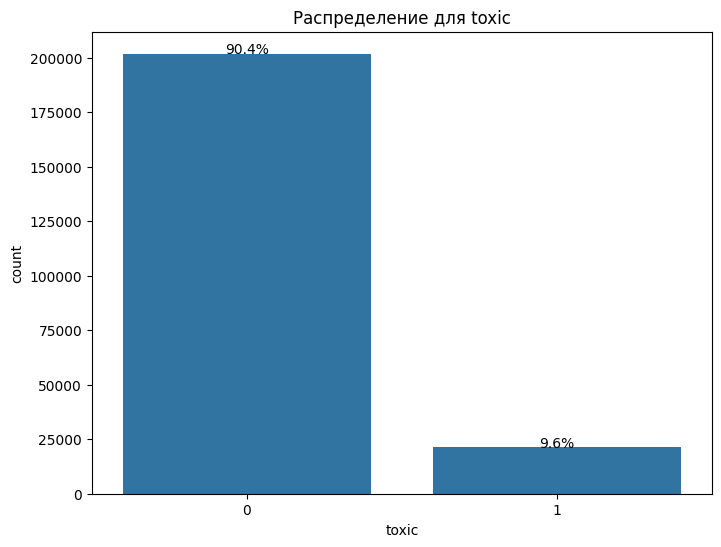




Судя по выводу df.info(), пропущенных значений в данных нет.

# Проверьте, есть ли дисбаланс классов (например, если 95% комментариев нетоксичные, а 5% токсичные).





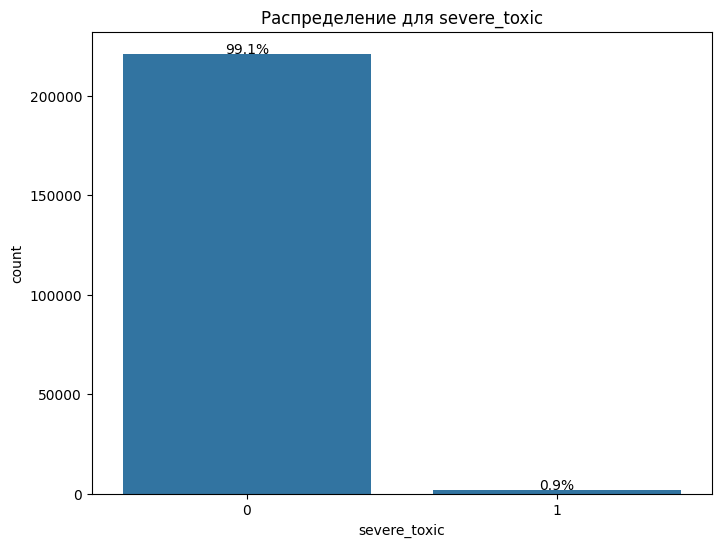
Распределение для toxic:

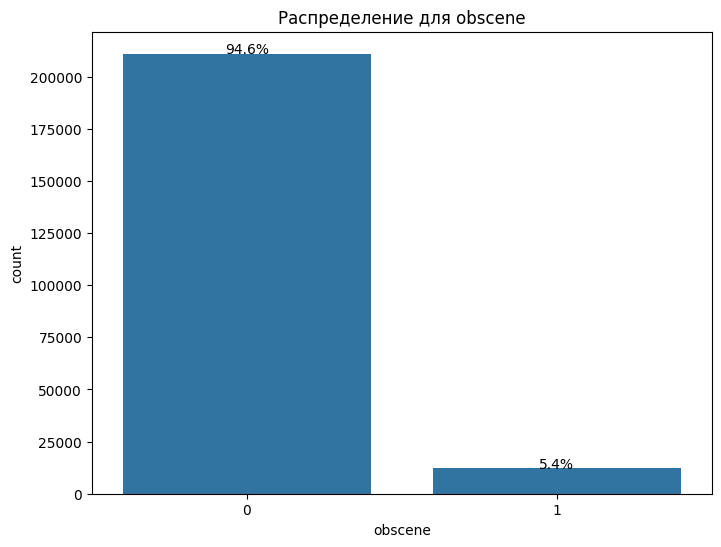
toxic

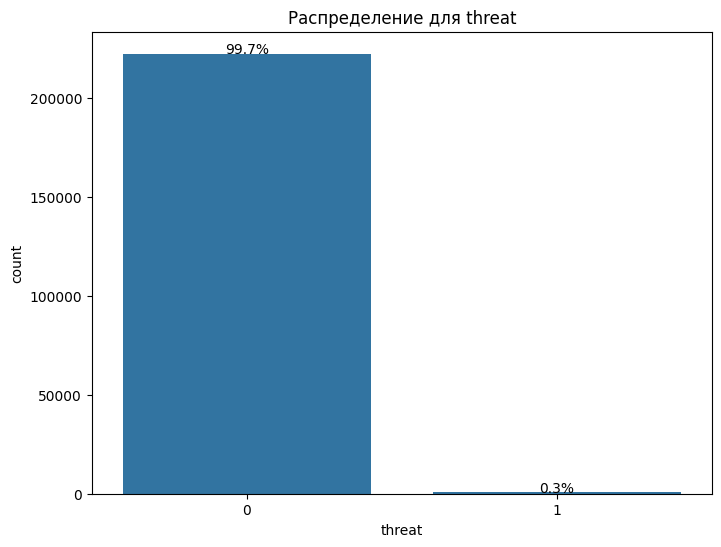
0 90.426938

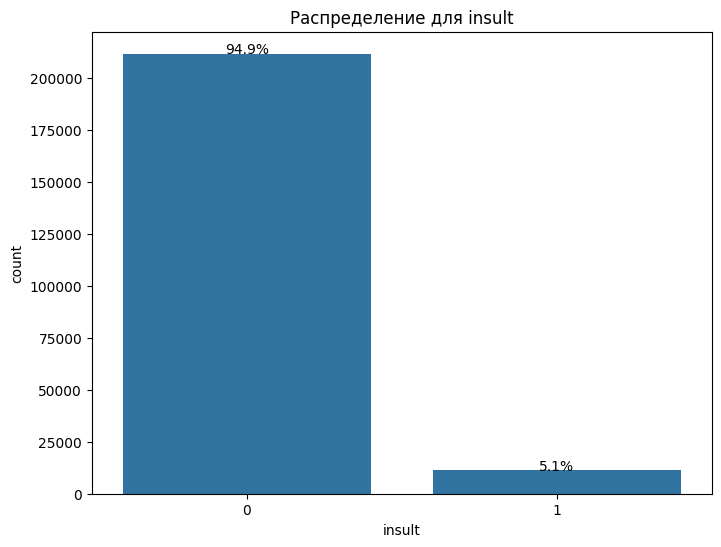
1 9.573062

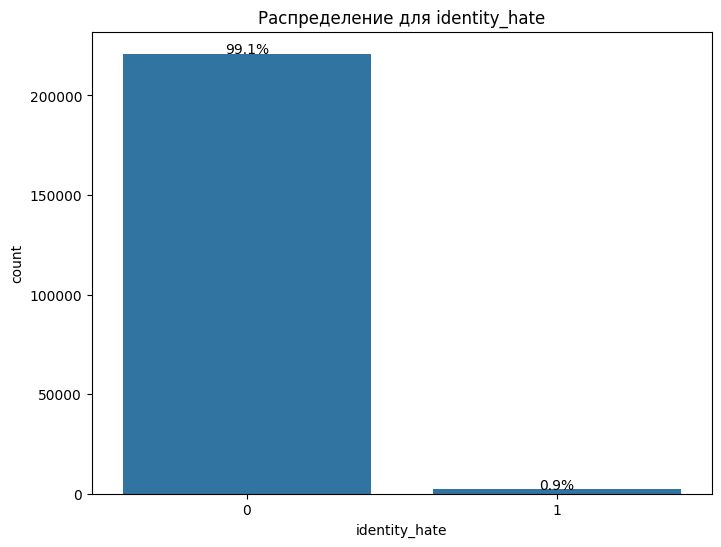
Name: proportion, dtype: float64



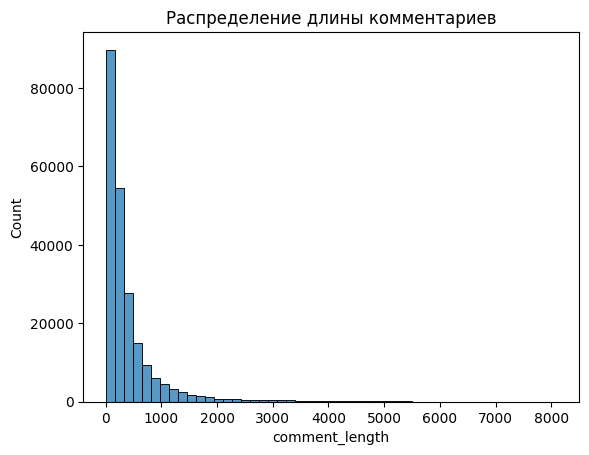


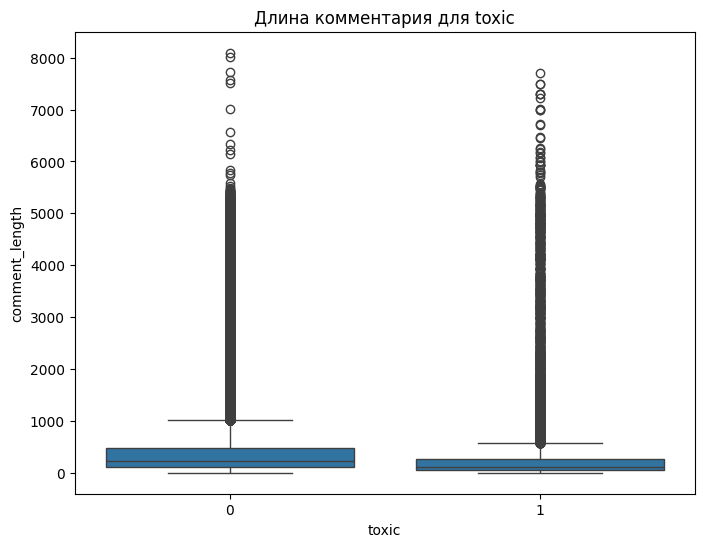


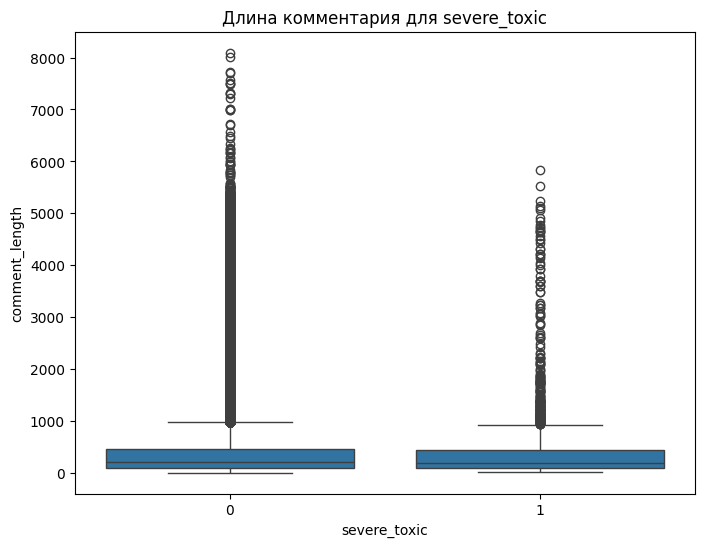


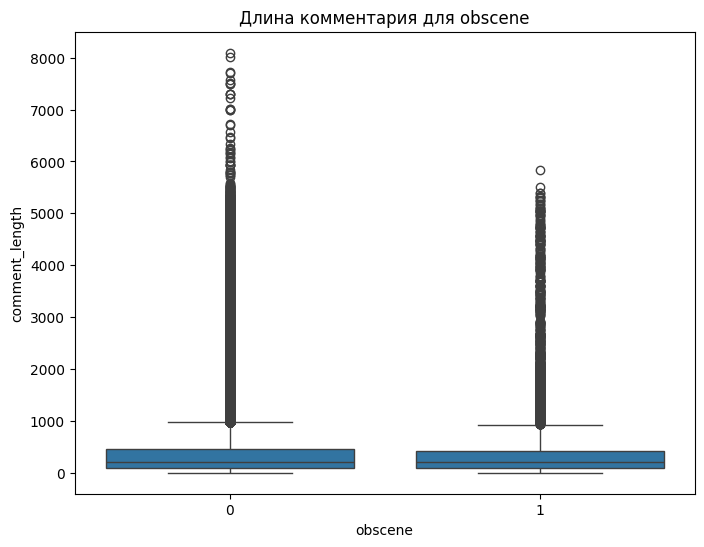


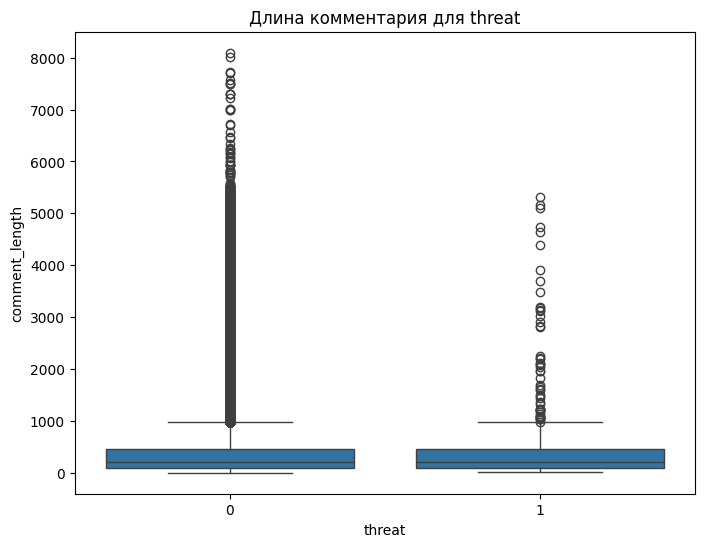
# Изучите длину комментариев

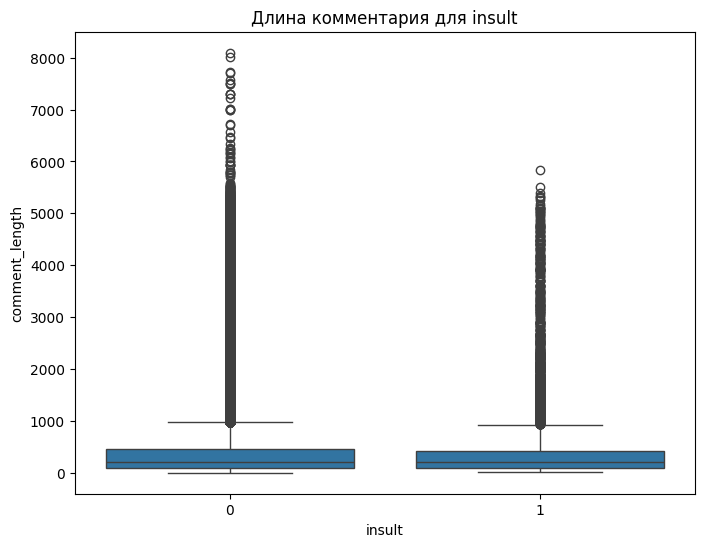
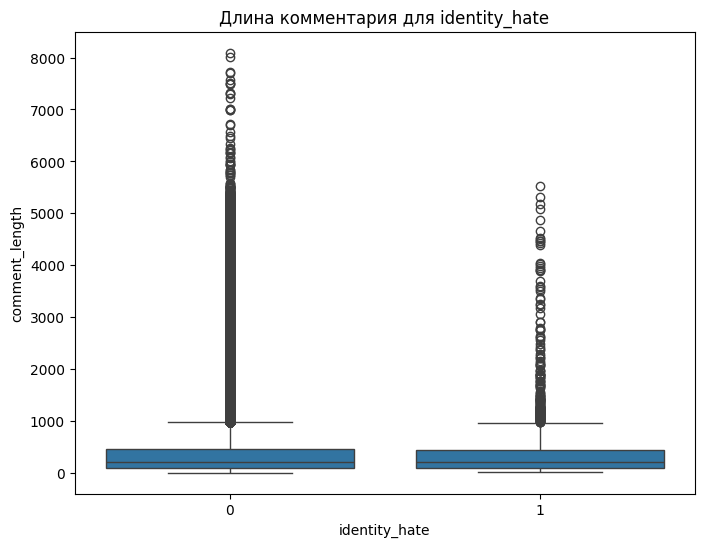




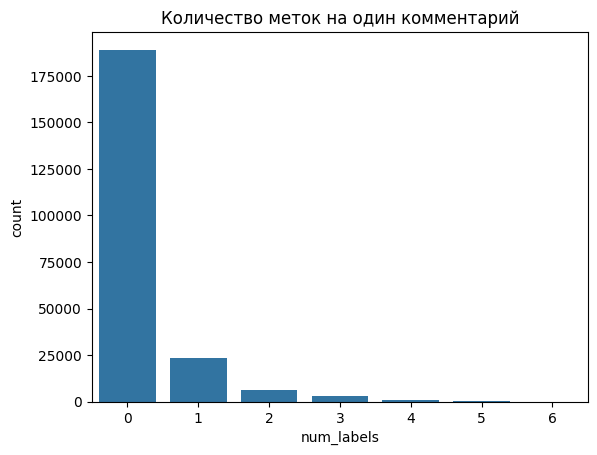


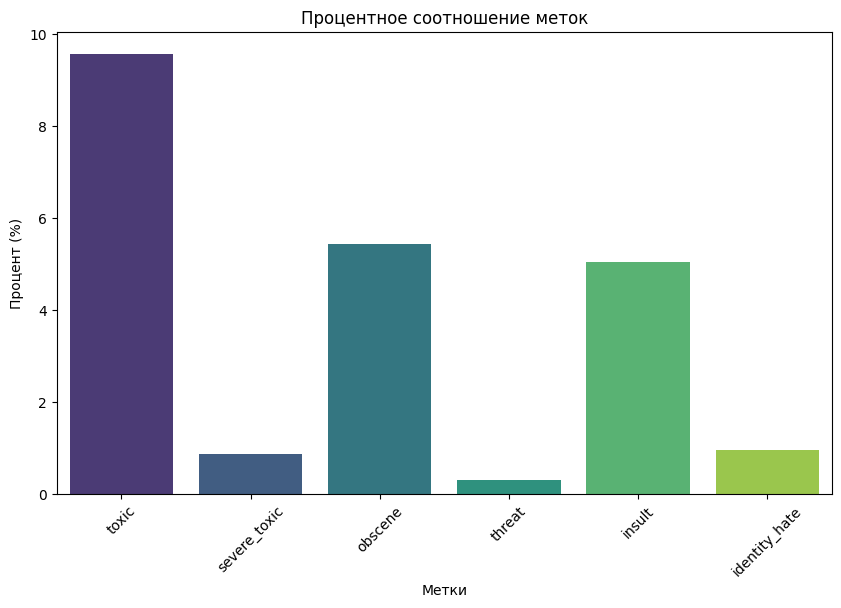


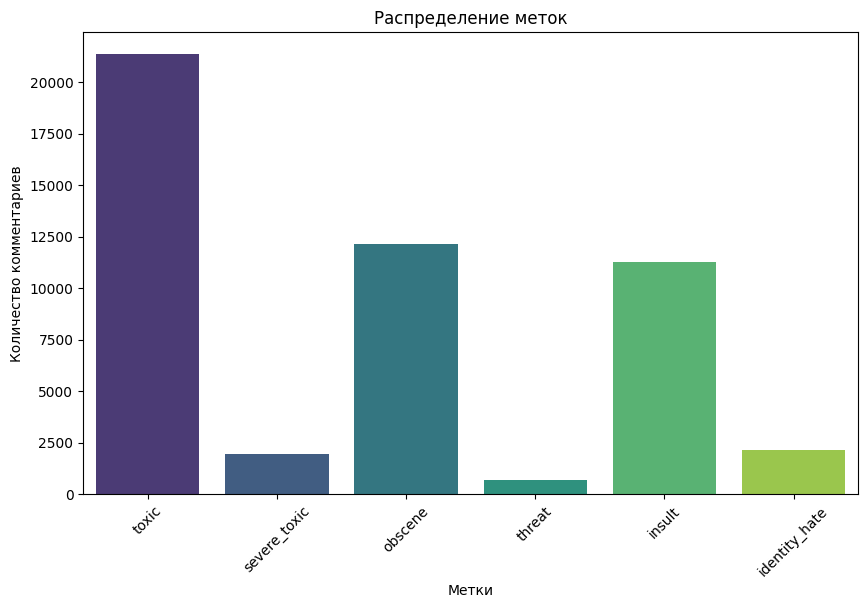


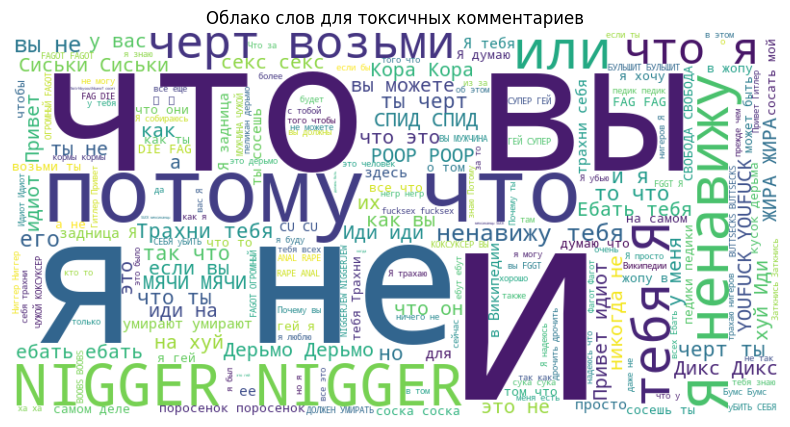
Обратите внимание, что комментарий может быть одновременно помечен несколькими метками (например, toxic и obscene).





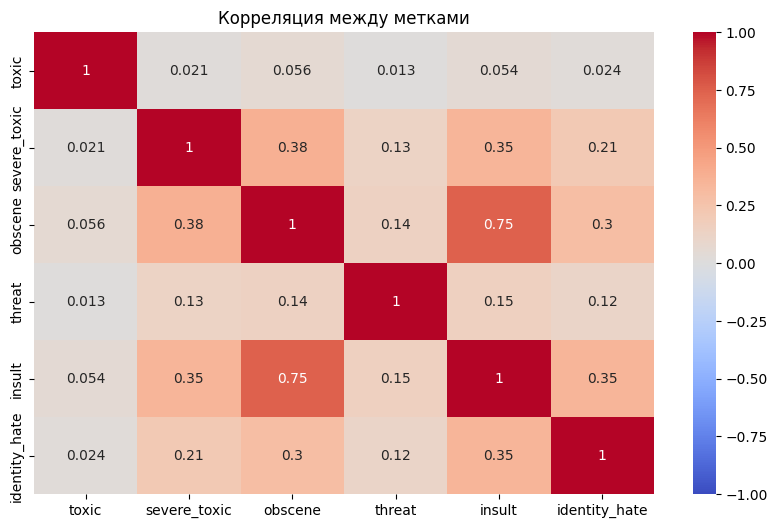


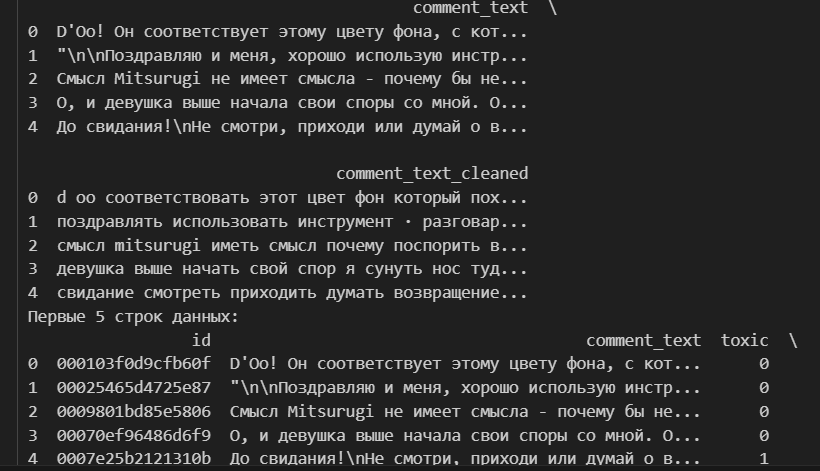
построения облака слов

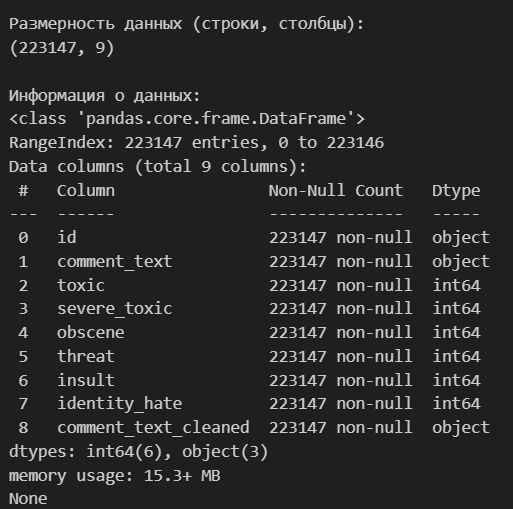




Корреляция между метками











C:\Users\bokar\Documents\GitHub\AI-Stream\detoxify\jigsaw\_data\jigsawtanslatedgoogle\ jigsaw-unintended-bias-train\_ru\_clean.csv

C:\Users\bokar\Documents\GitHub\AI-Stream\detoxify\jigsaw\_data\jigsaw-train-multilingual-coments-google-api\ jigsaw-toxic-comment-train-google-ru.csv

C:\Users\bokar\Documents\GitHub\AI-Stream\detoxify\jigsaw\_data\jigsaw-train-multilingual-coments-google-api\jigsaw-toxic-comment-train-google-ru-cleaned.csv

1. Предобратока данных
2. Удаляются столбцы Unnamed: 0.1 и Unnamed: 0 с помощью df.drop().

Текст приводится к нижнему регистру.

Удаляется пунктуация.

Удаляются стоп-слова с использованием библиотеки nltk.

Выполняется лемматизация с использованием библиотеки natasha.

Проверяются дубликаты по столбцу comment\_text.

1. Исследовательский анализ данных
2. Определение и обоснование метрик качества данных
3. Разработка базы данных для хранения данных
4. Оформление Сбора данных, Предобратоки данных и сохранения в базу данных для хранения данных в отдельный пайплайн для автоматизации
5. Оформление Дашборда
6. Оформление презентации/пояснительной записики о работе