Untitled

Пайплайн работы

Предобработка текста

Удаление нежелательных символов

Токенизация предложений

Приведение всех слов к нижнему регистру

Удаление стоп-слов

Лемматизация и стемминг

Удаление редко встречающихся слов

Векторизация текста

TF-IDF

1

Плюсы

- Простота и эффективность
- Учет важности слова
- Широкое применение

2

Минусы

- Не учитывает порядок слов в тексте и их семантическое отношение
- Чувствителен к шуму и опечаткам

Векторизация текста

Word2vec

1

Плюсы

- Учитывает семантическое отношение слов
- Широкое применение

2

Минусы

- Не учитывает контекст
- Большая требовательность к ресурсам

Векторизация текста

sBert

1

Плюсы

- Лучше работает на маленьких выборках
- Учет смысл и контекст в предложениях

2

Минусы

- Обучение занимает большое количество времени
- Большая требовательность к ресурсам

Выбор библиотеки машинного обучения

Логистическая регрессия

Плюсы

- Прост в реализации и интерпретации
- Эффективен для больших объемов данных

Минусы

- Признаки должны быть линейно независимыми, иначе результаты могут быть неточными

Может быть чувствителен к выбросам в данных

Catboost

Плюсы

- Автоматически обрабатывает категориальные признаки
- Имеет быстрый и эффективный алгоритм градиентного бустинга

Минусы

- Может быть медленным на больших объемах данных
- Может быть чувствителен к выбросам в данных

SVM

Плюсы

- Может обрабатывать большие объемы данных
- Может работать с высокоразмерными данными

Минусы

- Может быть чувствителен к выбросам в данных

Gradient Boosting из Scikit-learn

Плюсы

- Дает высокую точность прогнозирования
- Может использоваться для решения задач регрессии и классификации
- Работает относительно быстро

Минусы

- Может быть чувствителен к выбросам в данных

Точность алгоритма

	precision	recall	f1-score	support
-1	0.88	0.93	0.91	1046
0	0.65	0.45	0.53	267
1	0.84	0.88	0.86	624
accuracy			0.85	1937
macro avg	0.79	0.75	0.76	1937
weighted avg	0.84	0.85	0.84	1937

```
[[972  30  44]
 [ 86 119  62]
 [ 44  34 546]]
F1 score:  0.8373213009746296
ROC AUC score:  0.9313424692823705
```