Используя библиотеки sklearn и torch, в зависимости от датасета решите следующие задачи классификации (мультиклассификации) или регрессии. Графики рисуйте с помощью библиотеки matplotlib или любой другой удобной библиотеки (например, seaborn)

1. Нормализуйте текст

2. Разбейте его на трейн/тест. Если датасет слишком большой, то оставьте не больше 10к — 100к примеров

3. Обучите логистическую регрессию стохастическим градиентным спуском, используя tf-idf в качестве факторов. Редкие слова удалите. Проанализируйте, какие слова получили наибольший вес. Попробуйте лемматизировать слова. Нарисуйте loss на графике на трейне и на тесте (например, с помощью библиотеки matplotlib). Посчитайте метрику accuracy, F-1 (macro/micro для задачи мультиклассификации). Для задачи регрессии используйте MAE, MSE/RMSE.

4. Поэкспериментируйте с весом в L1 регуляризации. Сравните обученную новым способ лог. Регрессию с ранее обученным вариантом. Что лучше? Какие веса занулились? Нарисуйте графики. Посчитайте те же метрики.

5. Обучите нейронную сеть с помощью библиотеки torch с одним скрытым слоем, используя tf-idf над лемматизированными словами. Редкие слова удалите. Нарисуйте loss на трейн и тесте на графике. Сравните лоссы обученные с разными инициализациями: нулевая, xavier, he. В качестве функции активации используйте ReLU. Посчитайте те же метрики.

6. Зафиксируйте лучшее решение по ранее упомянутым метрикам.

7. В случае наличия дополнительных полезных нетекстовых данных для предсказания (площадь жилья, время отправки сообщения и т.п.) попробуйте заиспользовать их в подходе, показавшим наилучшее качество. Напишите, если таких данных в датасете не оказалось.

Датасеты

Egypt Real Estate Listings
https://disk.yandex.ru/d/kKpsSsg4vmd-vg
https://www.kaggle.com/datasets/hassankhaled21/egyptian-real-estate-listings

Задача: используя текст в description предсказать цену (price) (задача регрессии)

Financial News Market Events Dataset for NLP 2025
https://disk.yandex.ru/d/oz3Q060KmBSGnA
https://www.kaggle.com/datasets/pratyushpuri/financial-news-market-events-dataset-2025

предсказать impact level по описанию новости

IMDb 50K Cleaned Movie Reviews
https://disk.yandex.ru/d/k_l-yaY6HKiiwQ
https://www.kaggle.com/datasets/ibrahimqasimi/imdb-50k-cleaned-movie-reviews

По отзыву (review / cleaned review) предсказать сентимент (positive / negative)

Twitter Sentiment Analysis
https://disk.yandex.ru/d/dyJZThGPK2fYsA
https://www.kaggle.com/datasets/jp797498e/twitter-entity-sentiment-analysis

По тексту (tweet content) предсказать sentiment (positive / negative / neutral)

Phishing Email Dataset
https://disk.yandex.ru/d/XNMme-E55SwplA

https://www.kaggle.com/datasets/naserabdullahalam/phishing-email-dataset

По тексту из body
предсказываем label : 1 (фишинг), 0 — не фишинг

Russian Sentiment Dataset
https://disk.yandex.ru/d/tFJ0TivvzUt0KA
https://www.kaggle.com/datasets/mar1mba/russian-sentiment-dataset

По тексту предсказать sentiment