Разведывательный анализ данных

Был проведён разведывательный анализ данных для датасета, подготовленного для обучения LLM модели, в процессе которого было отмечено общее количество записей в датасете: 704 объекта вида: текст и класс экранной активности. Каждый класс содержит в общей сложности 176 примеров, что свидетельствует об отсутствии дисбаланса классов в датасете и как следствие позволит избежать переобучения модели на одном из классов.

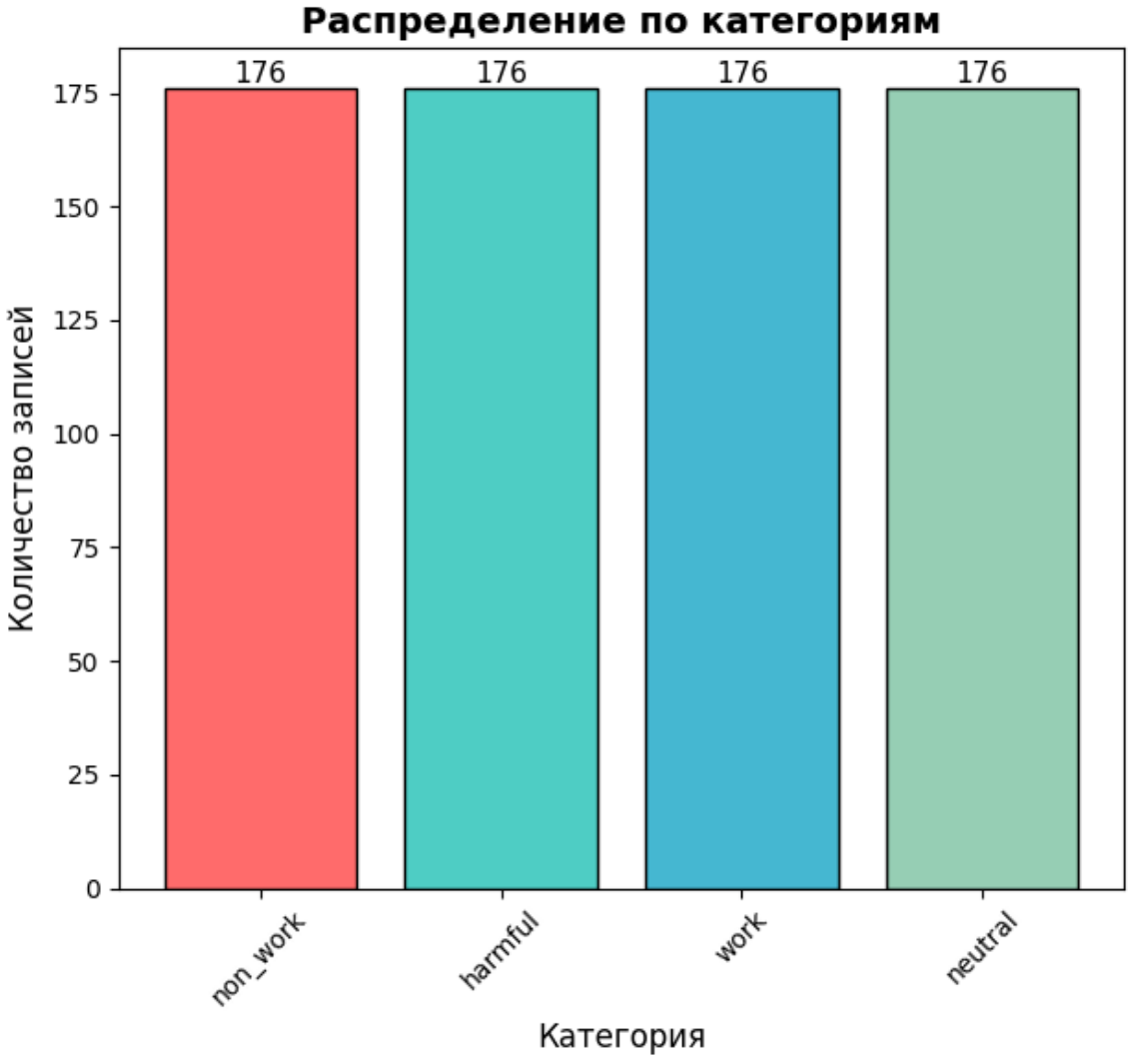

Рисунок 1 – Распределение объектов по классам

Также был проведен анализ распределения длины текстов и плотности их распределения в зависимости от класса. В качестве меры по приближению

средней длины текста в категориях его длина была ограничена до 500 символов. Таким образом были получены следующие результаты:

– Средняя длина текста – 396 символов, медиана – 471 символ;

– Средняя длина по категориям: harmful (вредоносные сайты) – 410 символов, non_work (нерабочая активность) – 400 символов, neutral (нейтральная / системная активность) – 396 символов, work (рабочая активность) – 379 символов.

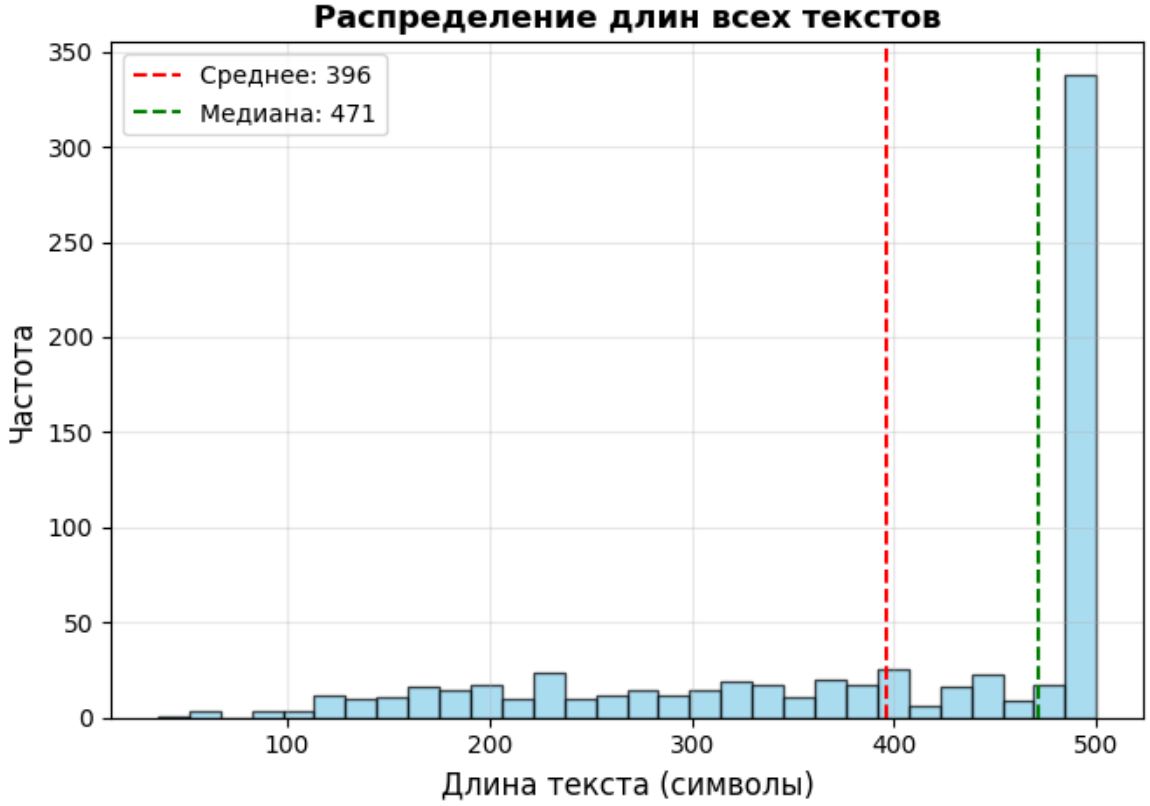

Рисунок 2 – Распределение по длине текста

Рисунок 3 – Плотность распределения классов по длине текста

Таким образом можно сделать вывод о том, что подготовленный датасет, собранный из текстов, распознанных со скриншотов из датасета website-screenshots и предварительно размеченных по видам экранной активности, не имеет дисбаланса в классах и может быть использован для обучения модели.

Построение и оценка модели

Для обучения модели исходный датасет был разделен на train, test и validation выборки в следующем соотношении:

– Train выборка – 492 примера;

– Validation выборка – 105 примеров;

– Test выборка – 107 примеров.

Модель rubert-tiny2 на таких данных показывает следующие результаты:

Таблица 1 — Динамика обучения модели

	Параметры обучения			
№ эпохи	Training accuracy	F1 Score	Training Loss	Validation Loss
1	0.24	0.20	1.38	1.39
2	0.31	0.25	1.36	1.37
3	0.43	0.40	1.29	1.31
4	0.48	0.49	1.14	1.28
5	0.56	0.56	0.94	1.16
6	0.58	0.58	0.79	1.08
7	0.66	0.66	0.51	0.98
8	0.64	0.64	0.42	0.95
9	0.68	0.67	0.32	0.89
10	0.70	0.70	0.23	0.89
11	0.70	0.69	0.16	0.92
12	0.71	0.71	0.15	0.95
13	0.69	0.68	0.13	0.97
14	0.72	0.72	0.10	0.98
15	0.72	0.72	0.10	0.99

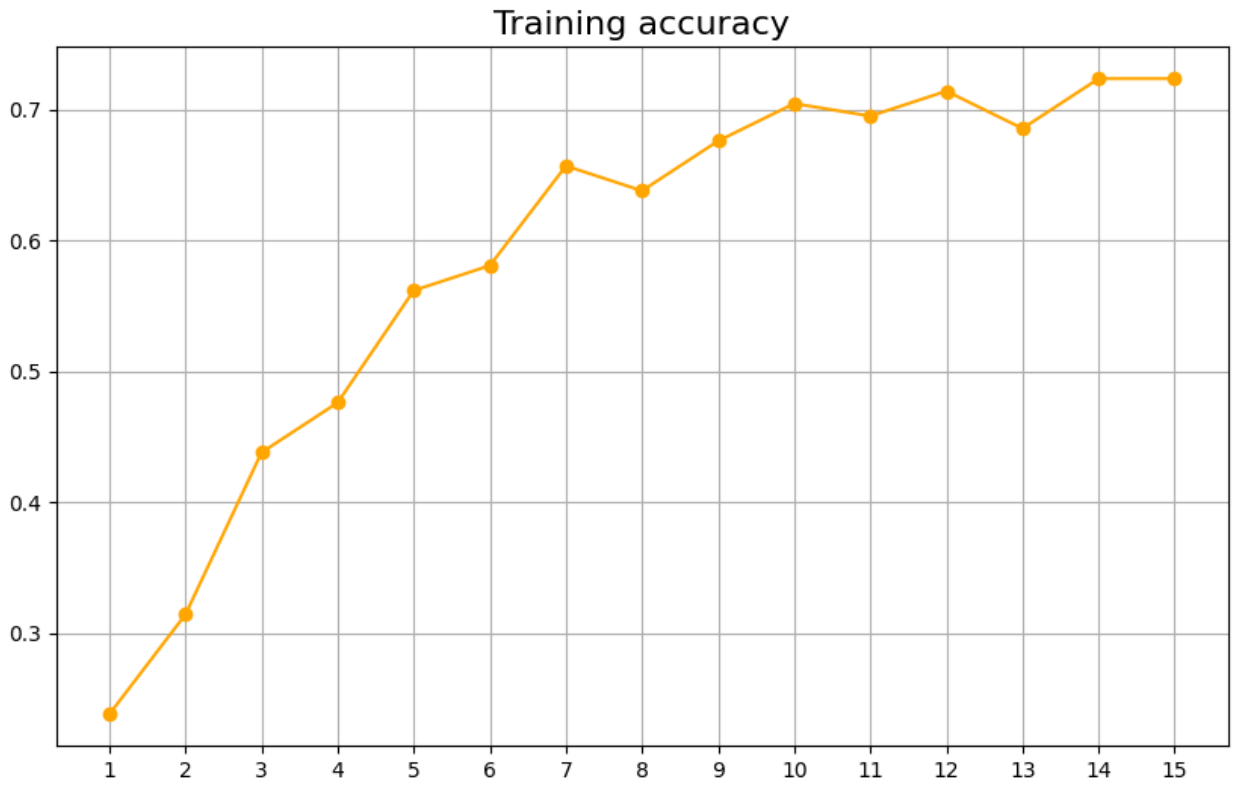

Рисунок 4 – Кривая обучения модели rubert-tiny2

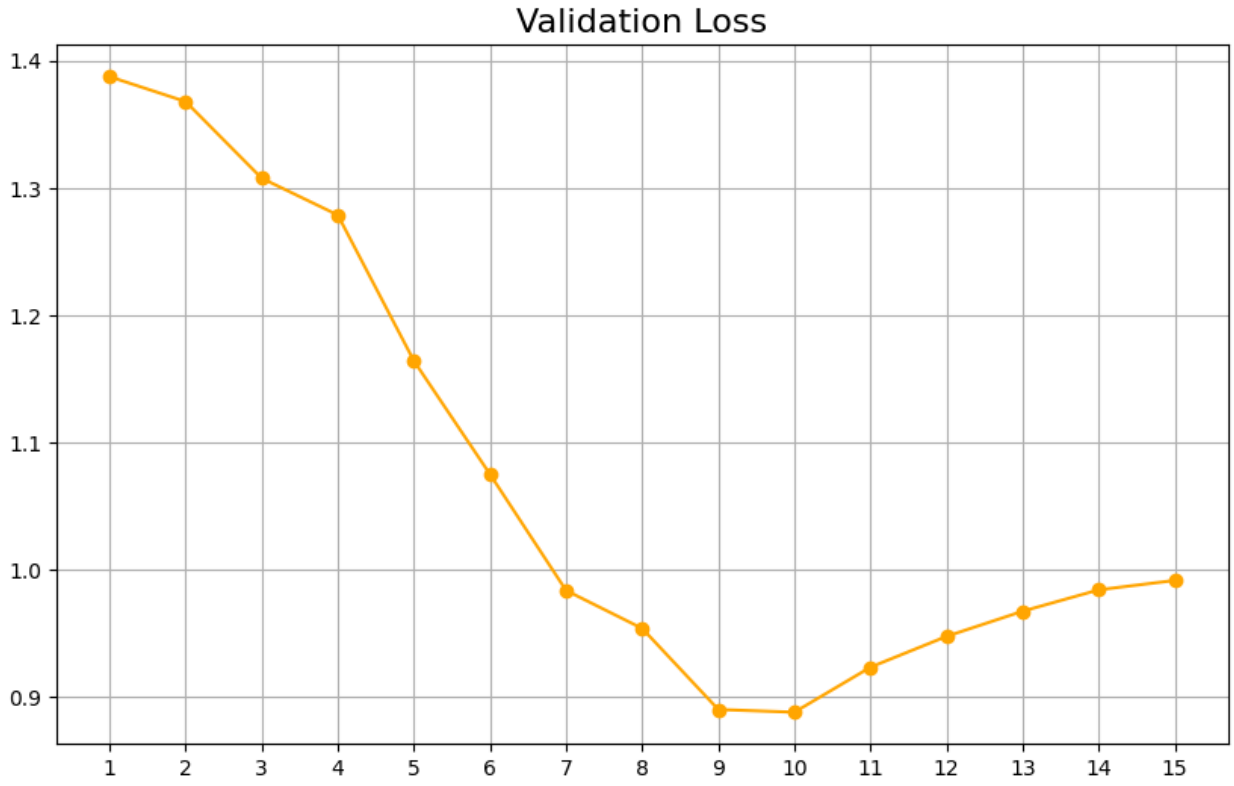

Рисунок 5 – Динамика Validation Loss обучения модели rubert-tiny2

Анализ полученных данных об обучении модели rubert-tiny2 показывает, что модель достаточно быстро обучается на первых эпохах, данные динамики обучения на поздних эпохах 11 – 15 свидетельствуют о наличии переобучения (validation loss начинает стабильно расти в то время, как training loss продолжает уменьшаться), что не так сильно сказывается на работе модели, которая при оценке на тестовых данных следующие результаты: accuracy – 0.84, F1 Score - 0.84.

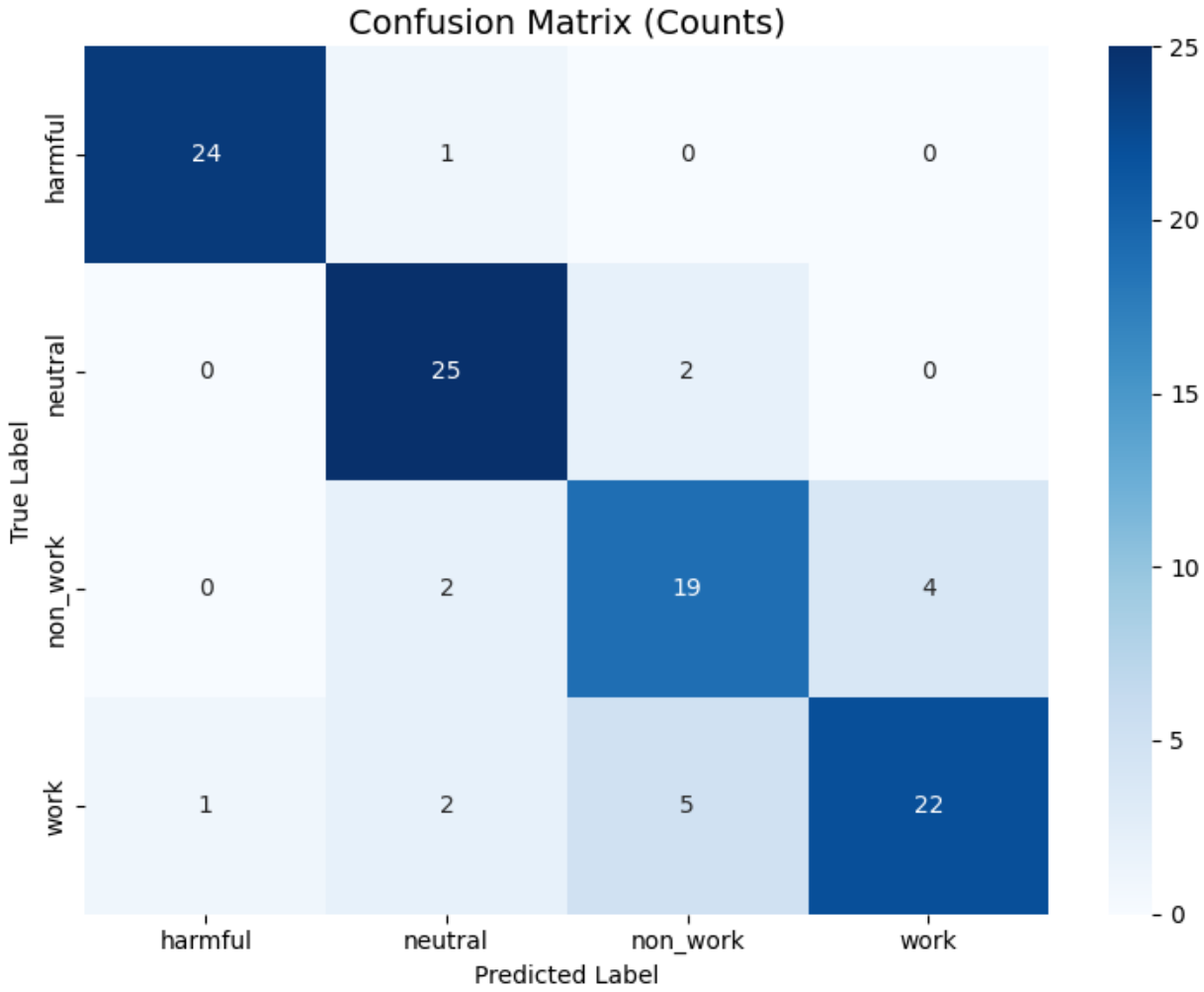

Рисунок 6 – Матрица ошибок модели rubert-tiny2

Построив матрицу ошибок, можно сделать вывод о том, что модель лучше справляется с определением классов harmful (вредоносные сайты): f1-score – 0.96, neutral (нейтральная / системная активность) f1-score – 0.88 и хуже

работает с классами work (рабочая активность) f1-score – 0.79, non_work f1-score – 0.75, в большинстве ошибочных случаев, путая их между собой. В дальнейшем, для улучшения качества работы модели можно уменьшить количество эпох обучения.

Таким образом, модель показывает хорошие результаты в классификации экранной активности при её оценке на тестовой выборке: accuracy – 0.84, F1 Score - 0.84, что свидетельствует о её готовности к работе в системе мониторинга.