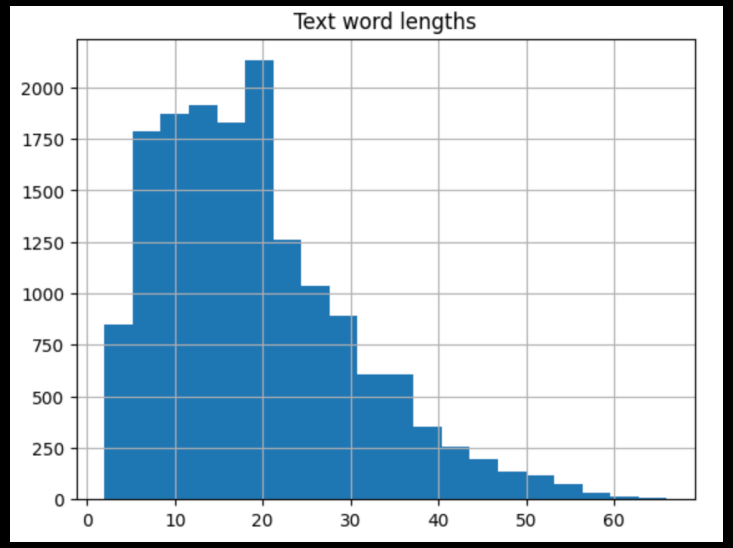
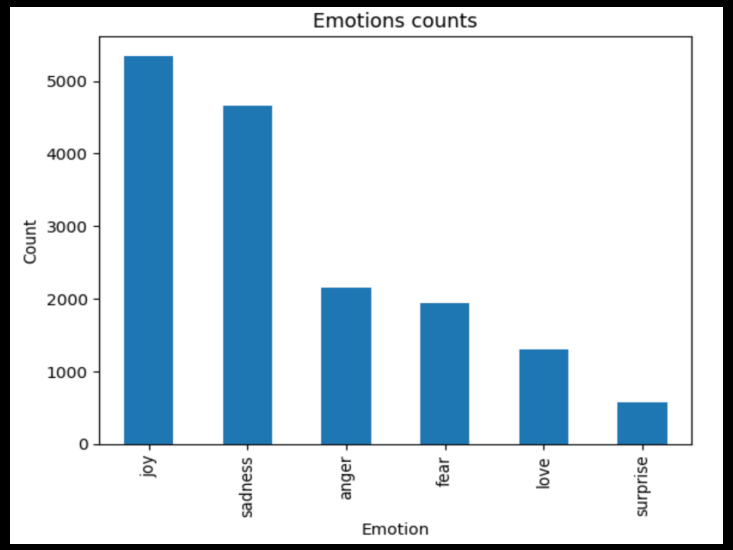
Технический отчет

Автор: Андрей Казначеев

Текстовый EDA





- Проверка на длину текстов
- Баланс классов
- Дубликаты
- Пропущенные значения (NaNs)
- Специальные символы
- Карточка датасета на HuggingFace

Выводы:

- Длина текстов не превышает 80 слов => не нужны специальные методы обработки длинных текстов
- Дисбаланс классов не катастрофический (всего один порядок), нужно попробовать учить с class-weights в loss-функции. В качестве метрики лучше использовать F1 score.
- Датасет монолингвальный и machine annotated => можно использовать модели english-only и трейн-вал-тест скоррелированы между собой

Мои решения

Классификация (с учителем)

Классическое обучение модели. Используется предобученная предсказывать замаскированный токен модель в качестве энкодера + классификационная голова, предсказывающая финальный класс. (Все варианты bert, bert-tiny, rubert-tiny2, roberta).

Zero-shot MLM

Не проводится обучение модели.
Используется предобученная предсказывать замаскированный токен модель. К тексту добавляется ”I feel [MASK] about it.” и сравнивается вероятность токенов, соответствующих целевой переменной. (т.е. какое слово сюда больше подходит – sad или joy или остальные 4 класса).

Zero-shot ChatGPT

Используется OpenAI API. Отправляем запросы модели. Даем ей затравку, что она бот, который определяет, какую эмоцию из списка передает текст и просим ее написать эмоцию одним словом.

Результаты

	Validation F1	Test F1
bert-tiny	0.863	0.825
rubert-tiny2	0.892	0.873
rubert-tiny2 ensemble	0.874	0.858
bert-base	0.918	0.873
bert-large	0.915	0.889
roberta-base	0.921	0.879
roberta-large	0.912	0.880
zero-shot bert-base	0.317	0.306
zero-shot ChatGPT	-	0.484

Для всех моделей перебирал параметры, результат представлен для лучшего набора (optimizer, lr, class-weights, param freezing, etc.). Для больших моделей class-weight немного улучшает перфоманс. Для маленьких – ухудшает.

Просматривая выгрузку ошибок ChatGPT кажется, что модель особо не ошибается, а кейсы очень пограничные. Она не «поправлена» на критерии разметки этого датасета, но кажется, что дает более человекоподобные результаты.

Zero-shot bert MLM подход неожиданно оказался плох.

Оптимизация

Для оптимизации возьмем победившую на тесте модель: bert-large.

	Validation F1 (2k texts)	Test F1 (2k texts)	Inference time	Size (MB)
No optimization (CPU)	0.915	0.889	1 min 52 sec	1362
No optimization (GPU)	0.915	0.889	~7 sec	1362
Onnx + quantization (CPU)	0.790	0.733	~1 min	335
Rubert-tiny2 (CPU)	0.892	0.873	~3 sec	117
Rubert-tiny2 (GPU)	0.892	0.873	~1 sec	117

Заключение

Лучшая модель – bert-large

Оптимальная модель – rubert-tiny2

Не стоит сбрасывать со счетов ChatGPT, так как просматривая выгрузку ошибок заметил, что даже человеку тяжело определить эмоции из многих текстов, например отделить joy и love, или sadness и anger. Судя по описанию датасета — его лейблы сгенерированы в автоматическом режиме, и быть уверенными в их 100% точности нельзя. Тем не менее из-за этого присутствует явная корреляция между трейном, валом и тестом, из-за чего можно научить модели обобщаться на тестовую выборку. Несмотря на это я бы больше доверял ответам ChatGPT, так как при выборочной проверке её ответы были более корректными, чем ground truth лейблы.