Классификация манипулятивных фрагментов в новостных текстах

Георгий Жаров
Научный руководитель: Константин Воронцов

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

27 июня 2023 г.

Задача поиска пропаганды

- Актуальность: В современном мире задача поиска и классификации пропаганды является важной проблемой в силу большого влияния пропаганды на человека.
- Цель: Создание модели классификации фрагментов пропаганды в русскоязычных новостных текстах.

Задача поиска пропаганды

Математическая постановка

Данные: (s, c), где s – это фрагмент текста, c – метка класса. Пусть S – пространство текстовых последовательностей s, t – токенизатор, а V – словарь всевозможных токенов предобученной модели. Тогда токенизатор работает следующим образом

$$t : S \to (V)^n,$$

где n – это фиксированная длина входного вектора предобученной модели.

$a(w)$ — рассматриваемая модель, w — параметры модели. P — пространство векторов из $\mathbb{R}^d$, таких что $\forall p \in P : \sum_{i=1}^{d} p_i = 1$.

Таким образом модель работает как

$$a : V \to P$$

Пусть теперь $\hat{c}$ — предсказание модели, оно получается следующим образом

$$\hat{c} = arg \max_i a((V)^n, w)$$

Для обучения используется кросс-энтропийная функция потерь $CE(y, p) = -\sum_{i=1}^{d} y_i \log p_i$, где $y_i \in \{0, 1\}$ – метка принадлежности к i-ому классу, p_i – вероятность принадлежности к i-ому классу.
Тогда вся задача представляется в виде следующей оптимизационной задачи

$$CE(y, a(t(s), w)) \to \min_w$$

Оценка качества и относительные метрики

Классические метрики для задачи классификации: accuracy, precision, recall, f1-метрика.
Предлагается использовать: относительная f1-метрика.

$$RelF1_i = F1(S_{ij}, S_{ik}),$$

где $S_{ij}, S_{ik}, -$ разметки i-ого текста j-м и k-м разметчиком

$$MAF1 = \frac{1}{M} \sum_{i=1}^{M} RelF1_i,$$

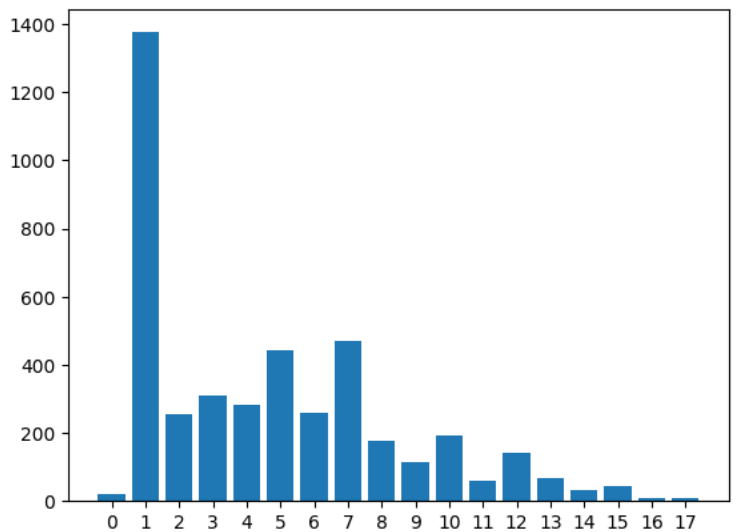
где $M -$ число текстов в датасете.

$$RF1 = \frac{F1(c, \hat{c})}{MAF1}$$

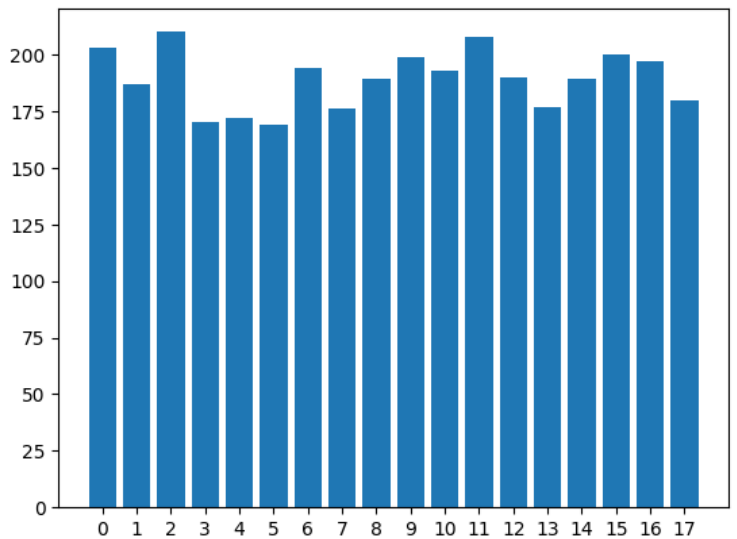
Данные

Проблема: сильный дисбаланс классов

Решение: равномерное сэмплирование обучающей выборки



а)



б)

Рис.: а) Распределение классов в исходной выборке
б) Распределение классов в сэмплированной обучающей выборке

Итоговая модель

Рис.: Общая схема используемой модели

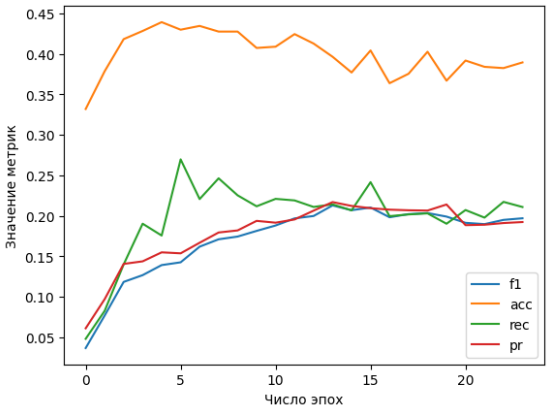
Финальная модель: токенизатор, предобученный энкодер (ruRoBERTa-large), линейный классификатор.

Вычислительный эксперимент

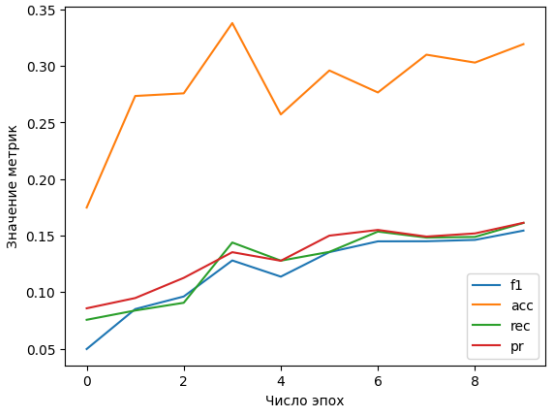
При обучении модели все слои предобученного энкодера кроме последнего замораживались. Обучение происходило с кросс-энтропийной функцией потерь. В качестве оптимизатора использовался Adam. Обучение происходило со следующими гиперпараметрами:

1. learning rate = 0.0001
2. label smoothing = 0.3
3. dropout = 0.3

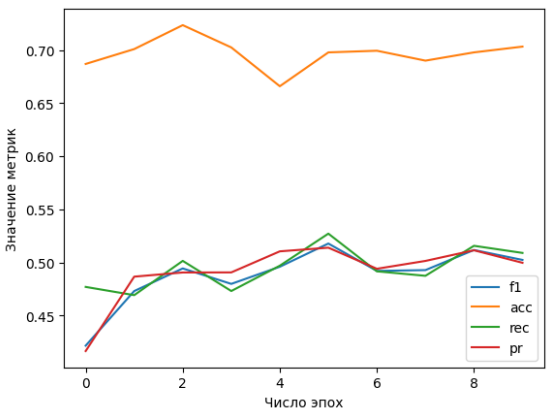
Результаты эксперимента



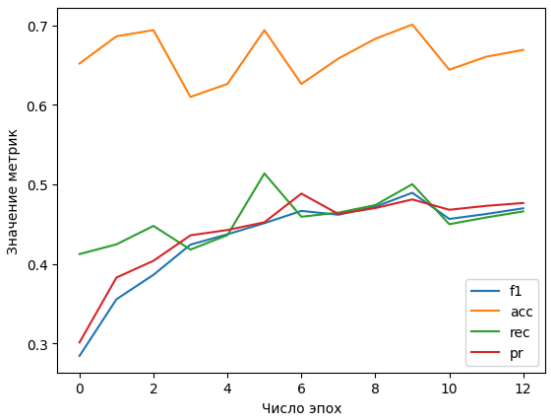
а)



б)



в)



г)

Рис.: а) cls18 б) cls18-context в) cls4 г) cls4-context

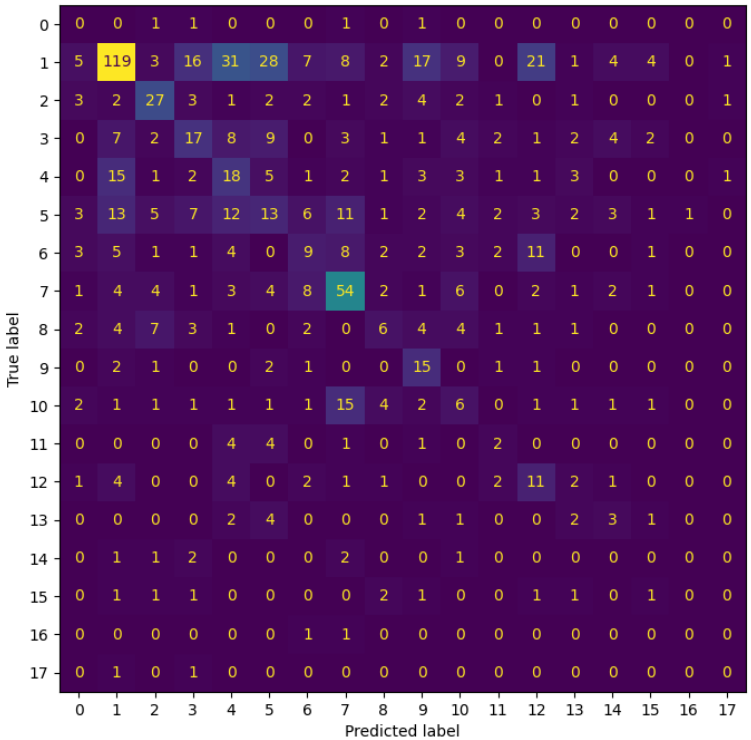
Метрики качества

В таблице ниже приведены основные значения метрик качества

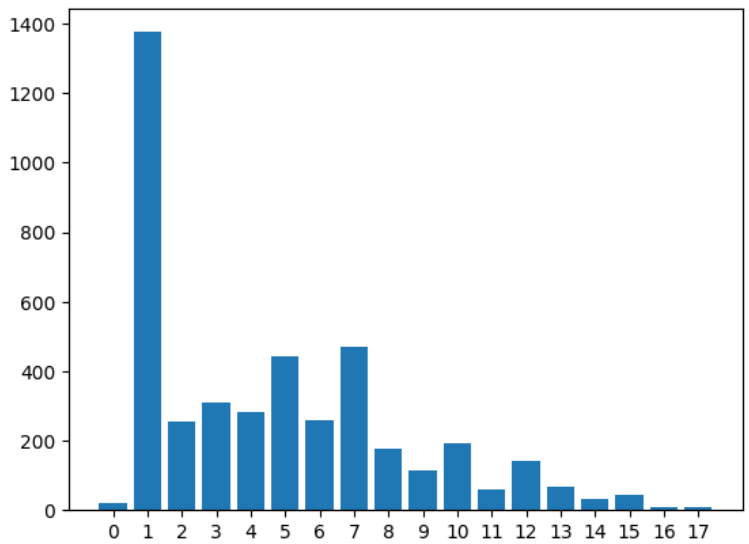
model	ACC	F1	RF1
cls18	0.353	0.249	0.508
cls18-context	0.357	0.166	0.339
cls4-context	0.620	0.443	-
cls4	0.676	0.509	-

Здесь cls18 – это модель, которая обучалась на данных, состоящих только из фрагментов пропаганды, cls18-context - аналогичная модель обучалась на фрагментах и их контексте, cls4-context и cls4 модели, классифицирующие фрагменты на 4 сгрупированных класса, обученные соответственно с контекстом и без.

Анализ ошибки модели



а)



б)

Рис.: а) Матрица ошибок модели
б) Распределение классов

На защиту выносится

- Построена и обучена базовая модель классификации фрагментов пропаганды в русскоязычных текстах.
- Реализован код для воспроизведения вычислительного эксперимента, поставленного в работе.