Автоматическое определение троллинга в текстах при помощи различных языковых моделей

Екатерина Соловкова, Яна Далевич, Светлана Жафярова

Recap: цель исследования

Проверить, как разные модели справляются с определением троллинга и его типа.

Гипотеза:

С выявлением скрытого (covert) троллинга модель будет справляться хуже, чем с определением явного (overt)

Recap: структура датасета

Title	Post	Troll comment	Response	TL	RL
If I'm not going to vaccinate myself, why?	Just heard that NC is considering giving portions of doses on-hand back to feds. If you've decided to not get jabbed, what's your reasoning?	I am glad you and a bunch of dumbs live in a nation that lauds your ignorance. covid is going to kill some of you idiots moving forward.	I got my shots, TYVM. I asked in general to attempt an antagonizing dialog with folks. Please try better, and remember you catch far more flies with honey than vinegar.	1	5
I think you guys complain too much	Everday I see posts like "there's too much damage", "too much mobility", ... I don't know. LoL has 140 champions and they all sit between 45-55% winrate, Riot got the one of the most popular games out there for 10+ years.	cringe post you can still delete this	Cringe comment You can still delete this	1	7

Table 2: Examples of collected Reddit posts, along with annotated strategies. TL: Troll strategy label; RL: Response strategy label. The number 1 of the TL column indicates overt troll. The numbers 5 and 7 of the RL column indicate *critique* and *reciprocate*, respectively.

Recap: стратегии троллинга

В таблице 1 приведено напоминание о том, что именно в [Lee et al. 2023] понимается под явным и скрытым троллингом

Overt strategies	**Aggress** is directly cursing or swearing others without any justification
	Shock is throwing an ill-disposed or prohibited topic that is avoided for political or religious reasons.
	Endanger is providing disinformation with the intent to harm others, and discovering this purpose by others.
Covert strategies	**Antipathize** is creating a sensitive discussion that evokes an emotional and proactive response in others.
	Hypocriticize is excessively expressing disapproval of others or pointing out faults to the extent that it feels intimidating to others.
	Digress is making a discussion to be derailed into irrelevant or toxic subjects.

Table 1: Types of troll behaviors (Hardaker, 2013)

Recap: задачи исследования

- Дополнить исходный датасет нейтральными примерами (без троллинга)
- Посмотреть, как различные модели справятся с выявлением троллинга в текстах в целом, сравнить результаты разных моделей
- Посмотреть, как модели справятся с определением типа троллинга и его выявлением в тексте в целом
- Проанализировать ошибки, чтобы проверить гипотезу

Выбранные модели

- RoBERTa: решение задачи классификации при помощи предсказания наиболее вероятного слова на месте маски в промпте
- Mistral: решение задачи классификации при помощи получения ответов от модели

Датасет

- Для решения задачи автоматического определения наличия троллинга в тексте нам пришлось дополнять исходный датасет нейтральными примерами (т.е., примерами без троллинга), поскольку исходный датасет содержал только троллинг.
- Мы отбирали примеры так же, как и в датасете с троллингом, состоящие из контекста - часто, вопроса - и ответа на этот вопрос.
- Количество набранных в итоге примеров - 600, но на практике мы использовали лишь 300 из них

Mask-prompting
классификация типа троллинга
(без учета контекста):

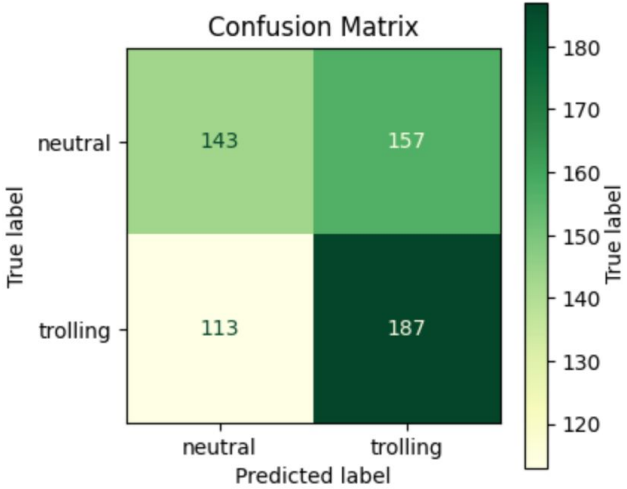
ROBERTA

Roberta: этапы работы

- Классификация через mask-промптинг (для каждого текста применялся шаблон с < mask >, затем модель предсказывала наиболее вероятные токены на ее месте)
- Сбор и категоризация предсказаний
- Оценка качества на тестовой выборке

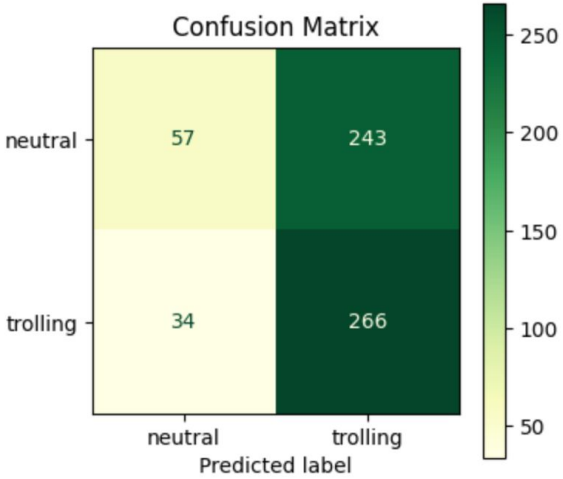
Замер качества и метрик по классам

0,5



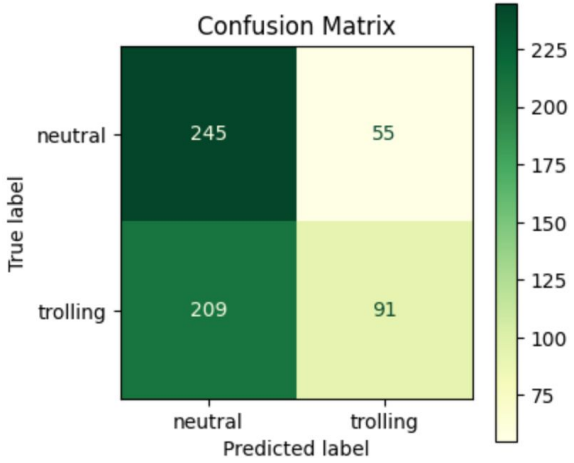
accuracy: 0.55
f1-score: neut - 0.51
troll - 0.58

0,7



accuracy: 0.54
f1-score: neut - 0.29
troll - 0.66

0,3



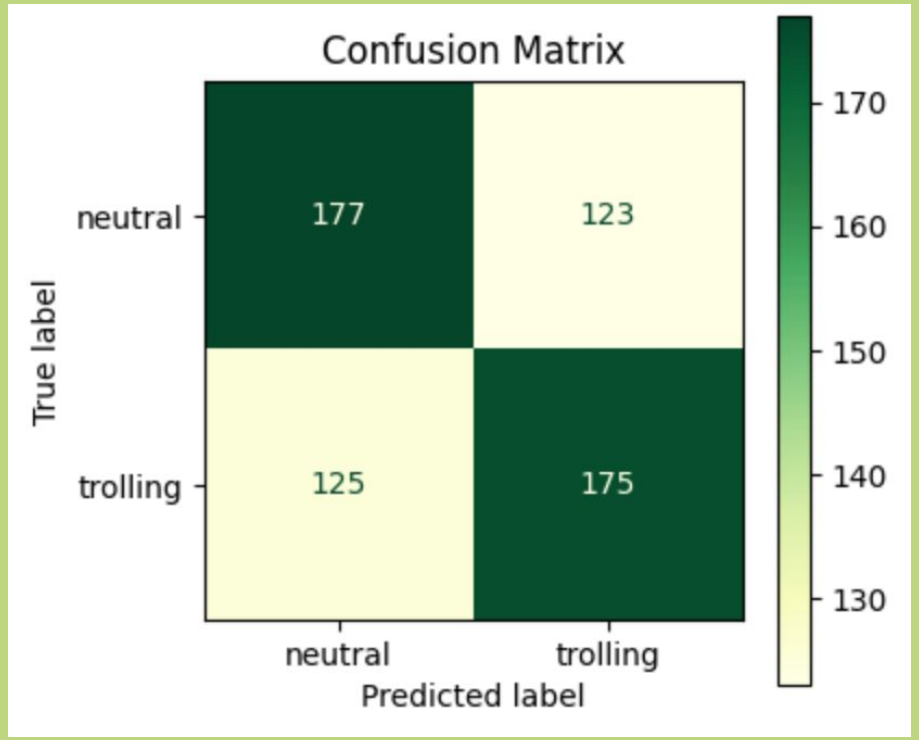
accuracy: 0.56
f1-score: neut - 0.65
troll - 0.41

Расширение списка слов за счет добавления предложений из датасета

Здесь нет смысла двигать границу классов. Результат хоть и улучшился, но все еще плохой.

граница классов - 0,5

accuracy: 0.59
f1-score: neut - 0.59
troll - 0.59



Roberta: выводы

- Для классификации overt/covert троллинга и в принципе его детекции без учета контекста не подходит метод mask-промптинга. Результат выходит чуть лучше случайного распределения.
- В теории, можно было бы подобрать больше промптов и выделить больше слов для категоризации, но вряд ли результат выйдет достаточно качественным.

Zero-shot детекция троллинга:

MISTRAL

Mistral: этапы работы

- получение ответов от модели (запрашиваемый формат ответа: 1/0 - наличие vs. отсутствие троллинга или 1/2 - явный vs. скрытый троллинг) с помощью промпта
- оценка качества на сформированной тестовой выборке
- анализ ошибочных ответов, чтобы подтвердить/опровергнуть первоначальную гипотезу
- эксперименты по подбору промпта с целью улучшить качество классификации
- сравнение результатов с другими моделями и методами.

Метрики качества

Приведем оценку качества ответов модели по классам:

	precision	recall	f1-score	support
class neut	0.76	0.91	0.83	300
class troll	0.89	0.71	0.79	300
accuracy			0.81	600
macro avg	0.82	0.81	0.81	600
weighted avg	0.82	0.81	0.81	600

Анализ ошибочных ответов

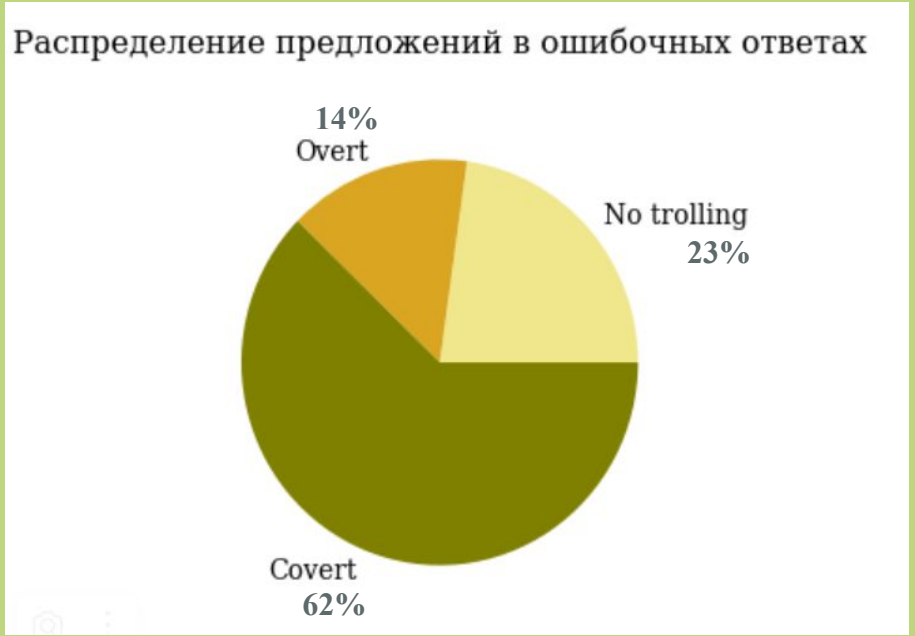
Гипотеза состояла в следующем:

> С выявлением скрытого (covert) троллинга модель будет справляться хуже, чем с определением явного

Посчитаем процентное содержание предложений без троллинга, с открытым и скрытым троллингом в ошибочных ответах модели, чтобы посмотреть, подтвердится ли гипотеза

Распределение предложений в ошибочных ответах

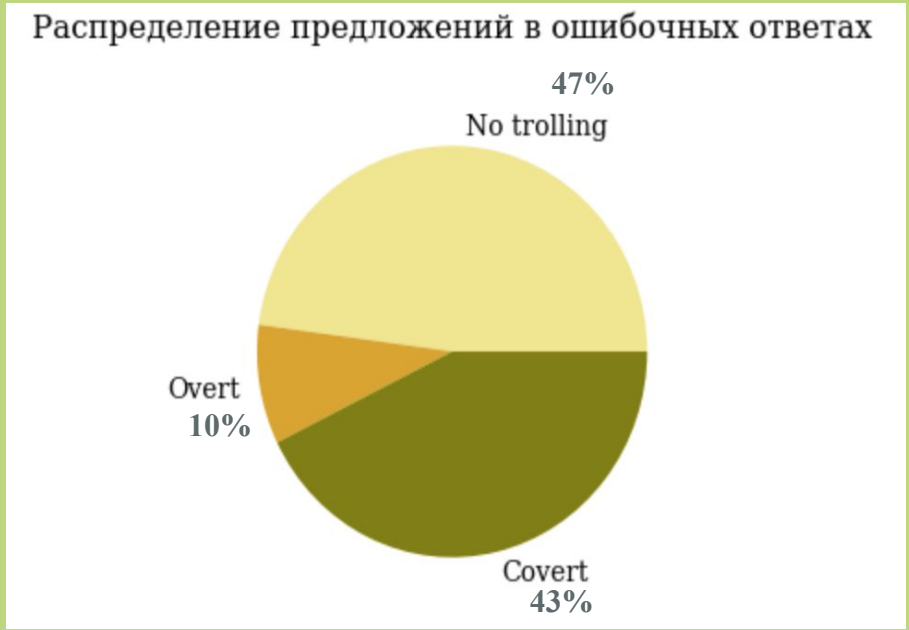
Как видно, гипотеза подтвердилась: среди ошибочных ответов действительно большинство примеров с covert типом троллинга



accuracy - 81%

Попытка улучшить ситуацию с закрытым троллингом

Качество улучшилось и процент скрытого троллинга в ошибках снизился (на 20%).



accuracy - 84%

Zero-shot классификация типа троллинга (overt/covert):

MISTRAL

Метрики качества

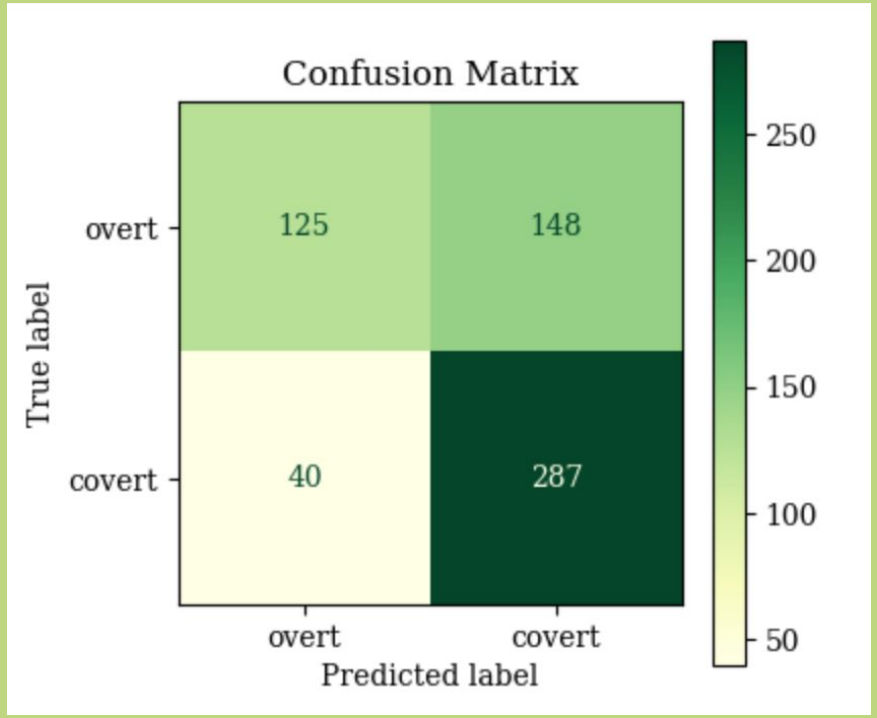
Приведем качество ответов модели по классам:

	precision	recall	f1-score	support
class overt	0.76	0.46	0.57	273
class covert	0.66	0.88	0.75	327
accuracy			0.69	600
macro avg	0.71	0.67	0.66	600
weighted avg	0.70	0.69	0.67	600

Видим, что имеет место заметное отличие качества по классам. Для наглядности выведем матрицу ошибок и посмотрим на распределение

Анализ ошибочных ответов

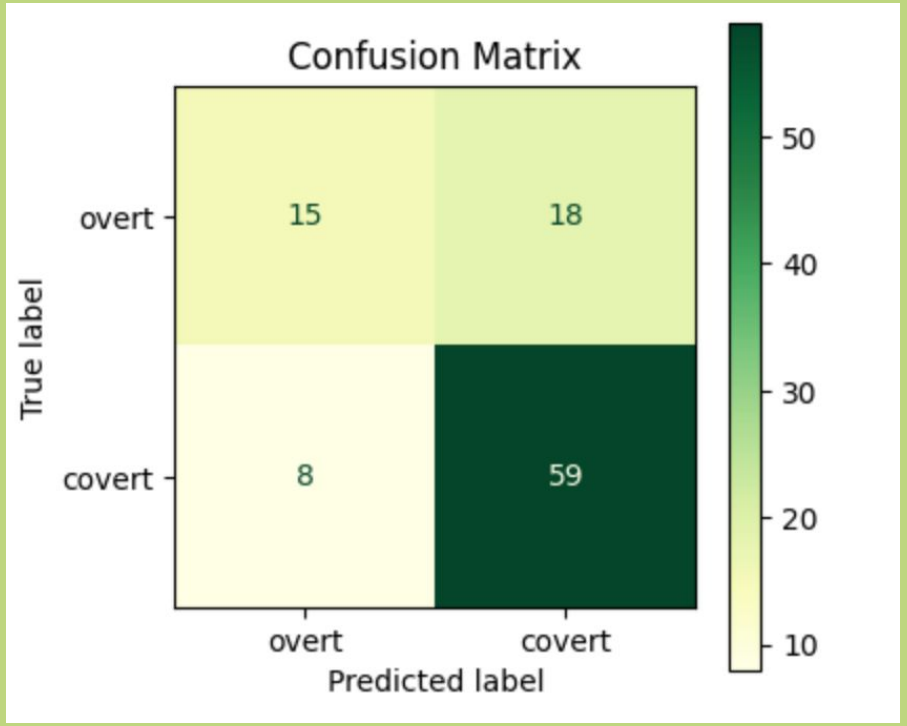
В данном случае наша первоначальная гипотеза не подтверждается: чаще ошибочно классифицируется открытый троллинг. Возможно, это влияние определений типов троллинга, данных в промпте.



accuracy - 69%

Попытка улучшить качество

К сожалению, улучшить классификацию открытого троллинга с помощью разъяснений в промпте не удалось: просмотрев несколько матриц ошибок, мы увидели, что процентное соотношение ошибок остается примерно одинаковым: в классе overt модель ошибается примерно в 50% случаев, в классе covert - примерно в 11-12%



Выводы

Подведем итоги экспериментов с Mistral

- с определением наличия/отсутствия троллинга в тексте Мистраль справляется довольно хорошо без дополнительного обучения -- максимальная точность, которой нам удалось достичь -- 84%
- с классификацией типа троллинга модель справляется чуть хуже -- максимальная полученная точность -- 69%
- в случае с определением наличия/отсутствия троллинга в тексте, наша гипотеза подтвердилась -- модель лучше распознает открытый троллинг и его в ошибочных ответах меньше всего
- однако в случае с классификацией типа троллинга гипотеза не подтвердилась: модель чаще ошибается в классификации открытого троллинга, а не скрытого. Такой исход, впрочем, кажется логичным -- в случае со скрытым троллингом в тексте нет явных маркеров негатива, а потому ошибочно отнести его к открытому классу троллинга сложнее, чем наоборот.

Выводы: итоговые метрики

	Task A			
	P	**R**	**wF1**	**MF1**
SVM	0.58	0.58	0.58	0.57
RF	0.60	0.59	0.54	0.52
BERT	**0.64**	0.63	0.63	0.63
RoBERTa	0.64	**0.64**	**0.64**	**0.64**

Таблица l: Результаты Lee et al. 2022

Task A - классификация типа троллинга, P - weighted precision, R - weighted recall; wFl - weighted Fl score; MFl -macro Fl score.

	Mistral		RoBERTa	
	Acc	MF1	Acc	MF1
Task 1	0.81	0.81	0.54	0.54
Task 1: best result	0.84	0.84	0.56	0.56
Task 2	0.69	**0.66**	-	-

Таблица 2: Наши результаты

Task l - наличие/отсутствие троллинга, Task 2 - тип троллинга; Acc - accuracy, MFl - macro Fl