Проектная работа

Автоматическое определение троллинга в текстах с помощью разных языковых моделей

24.02.2025

Состав команды

01

Екатерина Соловкова

- разработка плана работы
- генерирование идей

02

Яна Далевич

- поиск датасета
- структурирование инфомации

03

Светлана Жафярова

- создание презентации
- корректировка плана

Описание данных и процедура разметки

Типы троллинга

Overt trolling (открытый)

Агрессивные, явные оскорбления

Covert trolling (скрытый)

Более завуалированные, саркастические или манипулятивные комментарии

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)

01

- Аннотаторы читали комментарии и определяли их тип (overt/covert).

02

- Затем они анализировали ответы на trolling-комментарии и размечали их по одной из семи стратегий реагирования (например, вызов, игнорирование, насмешка и др.).

Пример разметки

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.

Обзор предыдущих исследований

На выбранном датасете проводилось исследование [Lee et al. 2022]. Какие модели использовались?

1. Классические ML-алгоритмы (SVM, Random Forest)
2. Глубокие нейросети (LSTM, CNN)
3. Предобученные трансформеры (BERT, RoBERTa, GPT)

Обзор предыдущих исследований

После составления и разметки датасета авторы проводили 3 эксперимента на указанных выше языковых моделях.

- задание A: автоматическая классификация trolling–комментариев по типам троллинга (overt/covert)
- задание B: автоматическая классификация ответных комментариев по типам стратегии ответа
- задание C: генерация ответа на троллинг с меткой стратегии реагирования. При оценивании результатов использовалась как автоматическая оценка (ROUGE-L Fl score, BLEU-l, METEOR, и BERTScore), так и человеческая оценка.

Обзор предыдущих исследований

Результаты для каждой модели и метрики оценивания можно увидеть в таблицах:

	Task C			
	R-L	**B-1**	**M**	**BS**
GPT-2	0.04	0.04	0.08	0.29
BART	0.08	0.08	**0.16**	**0.41**
DailoGPT-ELF22	0.06	0.14	0.04	0.35
GPT-2-ELF22	0.06	0.15	0.04	0.34
BART-ELF22	**0.10**	**0.15**	0.09	0.40

	Task A				Task B			
	P	**R**	**wF1**	**MF1**	**P**	**R**	**wF1**	**MF1**
SVM	0.58	0.58	0.58	0.57	0.32	0.34	0.33	0.19
RF	0.60	0.59	0.54	0.52	0.30	0.45	0.28	0.09
BERT	**0.64**	0.63	0.63	0.63	**0.48**	**0.47**	**0.47**	0.27
RoBERTa	0.64	**0.64**	**0.64**	**0.64**	0.46	0.42	0.42	**0.28**

Цель исследования

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

Гипотеза: “модели лучше справятся с открытым троллингом, так как он содержит явные маркеры агрессии и оскорблений”

Дальнейший план работы

01

Подготовка данных

- Очистка и предобработка текста.
- Добавление в датасет нейтральных текстов.

02

Выбор и обучение моделей

- Тестирование классических ML-алгоритмов и нейросетевых моделей
- Сравнение качества классификации.

03

Анализ результатов

- Метрики (Accuracy, F1-score, Precision, Recall).
- Ошибки моделей и их интерпретация.

04

Вывод и обсуждение

- Какие модели лучше справились с задачей?
- Какие трудности возникли?

Спасибо
за внимание !