



Автоматическое распознавание оскорбительных комментариев на русском языке в социальных сетях

Алла Горбунова, НИУ ВШЭ

Научный руководитель: Дарья Александровна Рыжова

Москва 2022

СОДЕРЖАНИЕ

01

Обзор

02

Данные

03

Эксперименты

04

Итоги



01

0Б30Р

ОПРЕДЕЛЕНИЕ ТОКСИЧНОСТИ

- токсичность по-разному определяют в разных работах, в том числе выделяют разные наборы классов внутри явления
- спорные классы: мат, спам, сообщения не по теме. В этой работе НЕ включаются в токсичность
- токсичность \approx оскорбления, угрозы и другие грубые негативные высказывания, направленные на человека или группу людей, выраженные явно или скрыто

ПОСТАНОВКА ЗАДАЧИ

- растет запрос на контроль за оскорбительными публикациями, и ручная модерация не справляется
- компании вводят автоматическую фильтрацию контента:
 - простой поиск по списку оскорбительных слов
 - алгоритмы машинного обучения
- алгоритмы охватывают больше материала, чем модераторы, но их легче обмануть
- чтобы избежать удаления комментариев, пользователи заменяют буквы в словах на 4ифры или зн@ки

ЦЕЛИ

- проверить влияние таких маскировок на качество работы сложных систем распознавания токсичности
- улучшить качество классификации с учетом этого фактора

СУЩЕСТВУЮЩИЕ ИССЛЕДОВАНИЯ

влияние маскировок на распознавание не исследовано

шесть работ про распознавание токсичности на русском (детали дальше)

три общих направления повышения качества:

- добавить данные
- улучшить архитектуру
- изменить подход к обучению

все найденные работы для русского не учитывают искаженные слова

очень краткие упоминания есть в работах для суржика и английского

ОБЗОР МОДЕЛЕЙ

RuBERT

[Smetanin, 2020]

F1-score 0,922

RuBERT

[Saitov, Derczynski, 2021]

F1-score 0,85

CNN

[Potapova, Gordeev, 2016]

accuracy 0,667

RuBERT

David Dale

общая метрика не указана

CNN

[Барсуков, 2021]

F1-score 0,872

Multinomial Bayes

[Smetainin, 2020]

F1-score 0,832

SVM

[Saitov, Derczynski, 2021]

F1-score 0,88

Random Forest

[Potapova, Gordeev, 2016]

accuracy 0,591

Logistic Regression

[Барсуков, 2021]

F1-score 0,781



02

ДАННЫЕ

ДАННЫЕ

доступны три корпуса токсичности на русском: RLTC, OK ML Cup и RUSSE Detox (пикабу, твиттер, одноклассники, двач)

источник новых данных — соцсеть ВКонтакте, группы Медуза, Дождь и Лентач

выбор обусловлен необходимостью использовать свежие данные, насыщенные как токсичностью, так и искажениями

данные собраны отдельно для двух этапов работы: 14.02.2022 и 27.04.2022



03

ЭКСПЕРИМЕНТЫ



ЭТАП 1: ПРОВЕРКА ГИПОТЕЗЫ

ГИПОТЕЗА: искажения слов, применяемые пользователями с целью обмануть автоматический фильтр комментариев, влияют на качество распознавания токсичности

3000 комментариев → 126 пар искаженных
+ исправленных вручную

проверяем все 6 моделей на двух датасетах:

- лучше на искаженных — гипотеза верна, маскировки скрывают токсичность
- лучше на исправленных — гипотеза верна, маскировки указывают на токсичность
- нет разницы — гипотеза не верна, маскировки не играют роли

	искаженные	неискаженные
токсичные	90	471
нетоксичные	36	2403

ЭТАП 1: ПРОВЕРКА ГИПОТЕЗЫ

гипотеза подтвердилась: все модели, кроме одной, работают лучше на исправленных данных

единственная модель, показавшая ухудшение (SVM), опирается на высокоуровневые признаки, в т. ч. тональность

на оригинальных данных больше FN, чем FP ошибок, исправление заставляет модели чаще отвечать «да» → FN снижается сильнее, чем растет FP

модель	F-score на искаженных	F-score на исправленных	
ruBERT, Smetanin	0,811	0,848	+ 0,037
ruBERT, Dale	0,662	0,745	+ 0,123
CNN, Барсуков	0,813	0,819	+ 0,006
MNB, Smetanin	0,578	0,679	+ 0,101
SVM, Saitov	0,810	0,800	- 0,010
LogReg, Барсуков	0,676	0,778	+ 0,102

ЭТАП 2: УЛУЧШЕНИЕ КЛАССИФИКАЦИИ

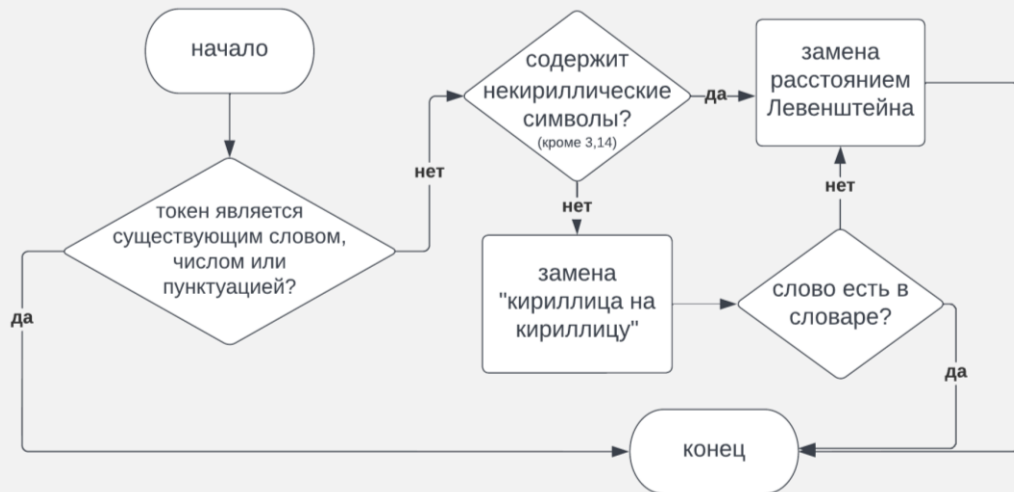
Два способа устранения уязвимости:

- 1) исправлять искажения по правилам
 - + не нужны дополнительные данные
 - + легко применять к любым классификаторам
 - много ошибок из-за особенностей орфографии в Интернете
 - не справится с искажениями, которых нет в правилах
- 2) обучать модель так, чтобы использовать искажения как признак
 - + не требует эмпирических правил
 - + потенциально может обобщать на новые типы искажений
 - нужно много данных и технических ресурсов

ИСПРАВЛЕНИЕ ИСКАЖЕНИЙ

49% искажений — замена на кириллические или латинские буквы, остальное — на цифры и символы (Прил. 4)

во всем тексте латиница заменяется на кириллицу по правилам (Прил. 5)



примеры правил

символ	замена
y	у
k	к
e	е
йо	е
цц	тс
3,14	пи

ИСПРАВЛЕНИЕ ИСКАЖЕНИЙ

алгоритм часто добавлял лишние слова, при этом модели показали даже более высокий результат, чем на исправленных вручную данных

тоже за счет снижения FN при небольшом увеличении FP

SVM снова выбился из общего паттерна, что подтверждает устойчивость его признаков к маскировкам

модель	F-score на искаженных	F-score на исправленных	F-score на предобр.
ruBERT, Smetanin	0,811	0,848	0,863
ruBERT, Dale	0,662	0,745	0,827
CNN, Барсуков	0,813	0,819	0,845
MNB, Smetanin	0,578	0,679	0,743
SVM, Saitov	0,810	0,800	0,810
LogReg, Барсуков	0,676	0,778	0,789

МНОГОЗАДАЧНОЕ ОБУЧЕНИЕ

идея опирается на эксперимент команды ВК с распознаванием protected identities

модель ruBERT от DeepPavlov

два подхода — добавление данных и добавление второй вспомогательной задачи:

- токсичность ВК без маскировок
- токсичность ВК + маскировки ВК
- токсичность микс без маскировок
- токсичность микс + маскировки ВК

используемые данные:

1. маскировки ВК: 1800 текстов, 50% класс с маскировками
2. токсичность ВК: 2897 текстов, 19% класс токсичности
3. токсичность микс: 8672 текста, 50% класс токсичности, 4 источника:
 - 2897 из ВК
 - 2000 из OK ML Cup
 - 1775 из RUSSE Detox
 - 2000 из RLTC

МНОГОЗАДАЧНОЕ ОБУЧЕНИЕ

лучший результат — сочетание добавленных данных и второй задачи

по отдельности вторая задача эффективнее добавления данных в первую

первая модель низкого качества и не учитывается в анализе

среди остальных обратная закономерность: двухзадачные модели лучше справляются на неисправленных данных

модель	F-score на искаженных	F-score на исправленных	F-score задачи 2
vk data + one task	0,342	0,298	
vk data + two tasks	0,721	0,599	0,722
mixed data + one task	0,683	0,782	
mixed data + two tasks	0,785	0,777	0,715

КОМБИНАЦИЯ ПОДХОДОВ

комбинация предобработки и двухзадачного обучения не показывает лучший результат

автоматическое исправление дает большой эффект только для однозадачных моделей

подтверждается, что двухзадачные модели учитывают маскировки как признаки, и лишая их этого, мы не улучшаем классификацию

модель	F-score на искаженных	F-score на исправленных	F-score на предобр.
vk data + one task	0,342	0,298	0,347
vk data + two tasks	0,721	0,599	0,617
mixed data + one task	0,683	0,782	0,824
mixed data + two tasks	0,785	0,777	0,788



04


ИТОГИ

ИТОГИ

- маскировки ухудшают распознавание токсичности — и их можно удалять из текста либо использовать как признак
- добавление лишних ругательных слов повышает качество распознавания
- и исправление маскировок, и двухзадачное обучение показали эффективность, но в комбинации они не усиливают друг друга
- однако в обучении добавление данных и добавление второй задачи лучше работают вместе
- в условиях ограниченных данных следует делить их на две задачи: такой подход дал больший прирост на меньших данных
- (квази)многозадачность может работать даже для простых архитектур

ДАЛЬНЕЙШАЯ РАБОТА

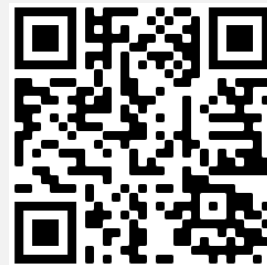


- максимальный порог расстояния Левенштейна
 - разные варианты правил замены
 - ансамбль из моделей и вариантов предобработки
 - другие типы задач
 - объем корпуса вспомогательной задачи равный основному корпусу
- 

ДОПОЛНИТЕЛЬНАЯ ИНФОРМАЦИЯ

обученные модели и все данные выложены в
репозиторий

корпус из 3000 комментариев ожидает публикации на
hatespeechdata.com



репозиторий

ИСТОЧНИКИ

МОДЕЛИ:

Барсуков Н. С. Выявление токсичного контента в русскоязычных текстах. Магистерская диссертация. Санкт-Петербург, 2021. 48 с.

Potapova R., Gordeev D. Detecting State of Aggression in Sentences Using CNN // SPECOM 2016. Budapest, Hungary: Springer Cham, 2016. P. 240–245.

Saitov K., Derczynski L. Abusive Language Recognition in Russian // Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing. Kyiv, Ukraine: Association for Computational Linguistics, 2021. P. 20–25.

Smetanin S. Toxic Comments Detection in Russian // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2020”. Moscow: The Russian State University for the Humanities, 2020. P. 1149–1159.

<https://huggingface.co/cointegrated/rubert-tiny-toxicity>

РАССТОЯНИЕ ЛЕВЕНШТЕЙНА:

Sood S. O., Antin J., Churchill E. F. Using Crowdsourcing to Improve Profanity Detection // Wisdom of the Crowd - Papers from the AAAI Spring Symposium. Palo Alto, California: The AAAI Press, 2012. P. 69–74.

МНОГОЗАДАЧНОЕ ОБУЧЕНИЕ:

Zueva N., Kabirova M., Kalaidin P. Reducing Unintended Identity Bias in Russian Hate Speech Detection // Proceedings of the Fourth Workshop on Online Abuse and Harms. Online: Association for Computational Linguistics, 2020. P. 65–69.

КАТАЛОГ HATE SPEECH DATA:

Vidgen B., Derczynski L. Directions in abusive language training data, a systematic review: Garbage in, garbage out // PLoS ONE. 2020. T. 15. № 12. P. 1–32.

**СПАСИБО
ЗА ВНИМАНИЕ !**