

Групповой проект

Студенты: Смилянский Александр, Бабич Кирилл

Наименование: The Hunt for Prohibited Content

Ссылка на соревнование: <https://www.kaggle.com/c/avito-prohibited-content/data>

1. Постановка задачи

Данные состоят преимущественно из товаров российского сегмента (русский текст). Задача состоит в классификации запрещённого контента по ряду признаков.

Avito является одной из крупнейших площадок для продажи вещей в России, преимущественно «из рук в руки». Однако, из-за масштаба возникает множество проблем с определением товаров для продажи. На Авито человек может сам определить категорию товара, составить описание и фотографию. На территории РФ продать можно далеко не всё, что угодно. Кроме того, ряд контента не имеет смысла в продаже или является «шуткой» пользователей. Примеры таких товаров: краденные вещи, ценные бумаги третьих лиц, спамерские базы, интеллектуальная собственность ГК РФ, а также нельзя продать воздух в районе, детей, и услуги вида «поеду вместо вас за границу». Первая группа запрещена Российским законодательством, вторая не имеет смысла в продаже. Обе группы запрещены правилами Авито.

Контроль за товарами на торговой площадке ведут модераторы. Для уменьшения затрат рабочей силы было решено составить алгоритм выявления некачественного (запрещённого правилами контента).

Задание: Составить модель, предсказывающую некачественное объявление о продаже на основе публикуемого материала.

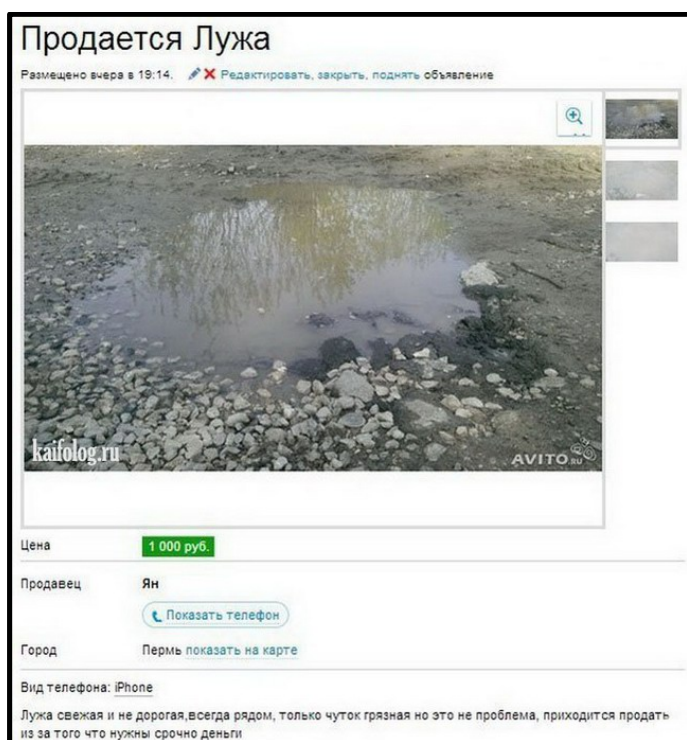


Рис. 1. Некачественный контент

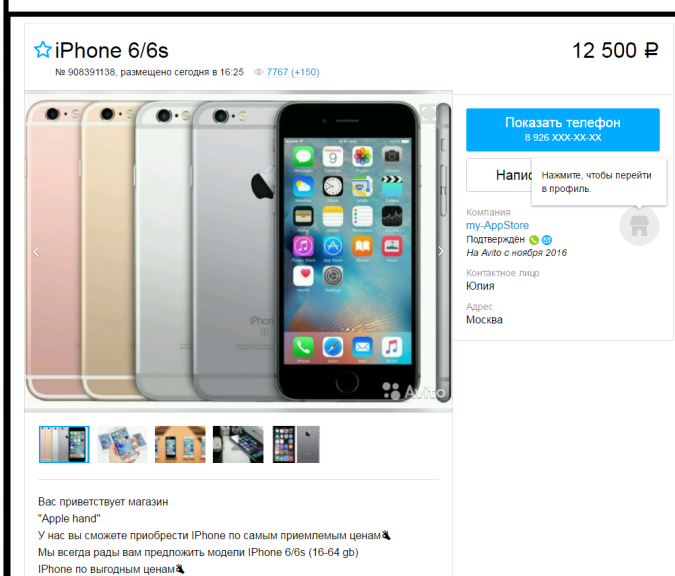


Рис. 2. Качественный контент

2.Описание данных

Данные в большинстве случаев состоят из русского текста. Файл закодирован в формате utf-8 и содержит tab разделители.

Данные содержат отдельные объявления и могли быть как заблокированы за недопустимый контент, так и никогда не подвергаться блокировке вовсе.



- **Itemid: int** – уникальный идентификатор
- **category: string** – категория первого уровня
- **subcategory: string** – категория второго уровня
- **title: string** – заголовок объявления
- **description: string** – Полный текст описания
- **attrs: JSON** – дополнительные параметры в формате JSON, которые соответствуют конкретному продукту, например конкретная модель автомобиля `{“car brand”:”bmw”, “car model”:”z1”}`
- **price: int** – финальная цена в Российских рублях
- **is_proved: boolean** – данный параметр доступен только в тренировочной выборке. Данный флаг выставляется только для заблокированного объявления по инициативе опытного модератора. Так как люди делают ошибки, то есть вероятность, что данное объявление было заблокировано неоправданно.
- **is_blocked: boolean** – **предсказываемая переменная**
- **phones_cnt: int** – количество телефонных номеров, найденных в описании. Если таковые имеются, то номер в описании заменяется на `@@PHONE@@`
- **emails_cnt: int** – количество электронных почт, найденных в описании. Если таковые имеются, то почта в описании заменяется на `@@EMAIL@@`
- **urls_cnt: int** – количество дополнительных ссылок, найденных в описании. Если таковые имеются, то ссылка в описании заменяется на `@@URL@@`
- **close_hours: float** – время жизни объявления на сайте Avito, Доступно только в тренировочных данных. Чем дольше объявление находилось на сайте и не было заблокировано, тем больше вероятность, что объявление не содержало запрещенного контента.

Category Subcategory
Все объявления в Москве > Транспорт > Автомобили с пробегом > BMW > X1
attrs

BMW X1, 2012

Title

Размещено 19 мая в 16:25. ✎ ✕ Редактировать, закрыть, поднять объявление



Цена 1 200 000 руб. [Онлайн калькулятор Каско](#)

Price

Продавец Владимир

[Показать телефон](#)

Город Москва [показать на карте](#)

BMW X1, 2012 г.
Пробег 30 000 - 34 999 км., 2.0 AT, бензин, полный привод, кроссовер, левый руль, цвет чёрный
2 литра, 184 л.с. Полный привод. 8 ступ. автомат, климат контроль, подогрев сидений, зимняя резина и так далее...
Не бита, не крашена.
Пробег 32 000км.
Машина обслуживается только у официального дилера.

Description

Номер объявления: 335309783

Рис. 3. Графическое представление данных

3. Анализ данных

Данные представлены выгрузкой из базы данных сервера и далеко не все посты были проверены администраторами. Посмотри на начальные данные, кол-во пропусков и категории некоторых данных.

Столбцов	3995803
Строчек	12

Содержат пропуски:

category	False
subcategory	False
title	True
description	True
attrs	True
price	False
is_proved	True
is_blocked	False
phones_cnt	False
emails_cnt	False
urls_cnt	False
close_hours	False

Пропуски существуют в колонках «Наименование», «Описание», «Атрибуты», «Подтверждено». «Подтверждено» - важный показатель при обучении на первый взгляд. Он приводится только для заблокированных объявлений и информирует, что модератор подтвердил блокировку.

Рассмотрим наш target столбец – is_blocked.

0	3720807
1	274996

Из всех объявлений, заблокированных 274996, это примерно 7.4% от всех объявлений. На этом и будет обучаться алгоритм.

Общая задача состоит из подпунктов:

1. Заполнить пропуски
 - a. Анализ пропусков
 - b. Выбор наилучшего алгоритма заполнения
2. Выбор подходящих моделей обучения, например:
 - a. Random Forest
 - b. Gradient Boosting
 - c. SGD
 - d. SVM
3. Настройка кроссвалидации
4. Выделить весомые признаки
5. Определиться в необходимости анализа текстовых признаков и его оценивании в модели
6. Определиться с необходимостью учёта is_proved
7. Обучение и настройка алгоритмов с лучшими параметрами
8. Улучшение модели

Анализ текстовых признаков и подбор лучшего алгоритма будут являться самыми сложными пунктами. По всем пунктам предоставить анализ и шаги к формированию конечного решения.