ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

**Московский институт электроники и математики
им. А.Н. Тихонова**

Студенты учебной группы МСКМ 181:
Чертенков В.И.
Пехтерев Д.О.
Горчавкина А.А.

Техническая документация по проектной работе

"Анализ отзывов о лекарственных препаратах в социальных медиа"

Руководитель проекта

Артамонов С.Ю.

Москва 2019

1 Существующие компании

На рынке услуг для анализа лекарственных препаратов уже сейчас существует ряд компаний, предлагающих своим клиентам получить обратную связь от потребителей о качестве их товаров и услуг, на основе имеющихся открытых источников в Интернете.

2 Многоклассовая классификация

Задача, в которой классов, на которые нужно разделить объекты, больше, чем 2. В отличии от бинарной классификации, где возможные значения классов +/- 1. Существует ряд методов, которые сразу решают задачу многоклассовой классификации, но в моем случае я свел задачу многоклассовой классификации к задаче бинарной классификации.

Были реализованы следующие методы:

1. СВМ – линейный классификатор. Просто реализуется, хорошо работает с большим объемом данных, любыми типами признаков (вещественные, категориальные и разреженные). 2. Случайный лес – строит композицию независимых друг от друга классификаторов (решающих деревьев). А после, путем голосования выбирается тот класс, к которому его отнесло большинство классификаторов 3. Градиентный бустинг – также строит композицию классификаторов, но использует взвешенное голосование, где каждый классификатор имеет собственный вес при голосовании. А также, в отличии от СЛ, где классификаторы строятся независимо друг от друга, в бустинге базовые алгоритмы строятся друг за другом, исправляя ошибки предыдущих алгоритмов.

Набор данных Представляет собой текстовые отзывы потребителей сервиса Amazon, извлеченные пользователем Julian McAuley, разделенных на 24 категории. Я проводил классификацию по 3 категориям из 24.

- Медицина и здоровье - Красота и уход - Товары для детей

Эти группы выбраны не случайно. Они относятся к близким предметным областям и используют схожу лексику, в отличии от более контрастных групп (например, если бы на ряду с медициной рассматривались электроника и кино).

Процесс обучения Из каждой из 3-х групп было взято одинаковое количество отзывов. Далее, данные разделы на 2 группы в соотношении 70/30. На 70% данных мы обучаем наши модели (т.н. обучающая выборка), а на оставшихся 30% мы оцениваем их качество (т.н. тестовая выборка).

3 Метрики качества

Существует несколько видов метрик качества, которые позволяют оценивать качество алгоритмов с разных сторон. Мы оценивали качество классификации тремя видами: 1. ДОЛЯ ПРАВИЛЬНЫХ ОТВЕТОВ (Accuracy). Показывает число верных ответов к общему числу отзывов в выборке. Это самая простая и самая понятная метрика, но не всегда отражает реальную эффективность алгоритма. 2. СРЕДНЕЕ ГАРМОНИЧЕСКОЕ (F2)..(,)3.$ROC-(AUC-ROC)$..,11,0.

4 Результаты многоклассовой классификации

Для всех трех методов были подобраны параметры, при которых каждый метод работает лучше всего. На рис.1 представлены показатели качества этих трех методов, а также параметры при которых достигаются эти показатели.

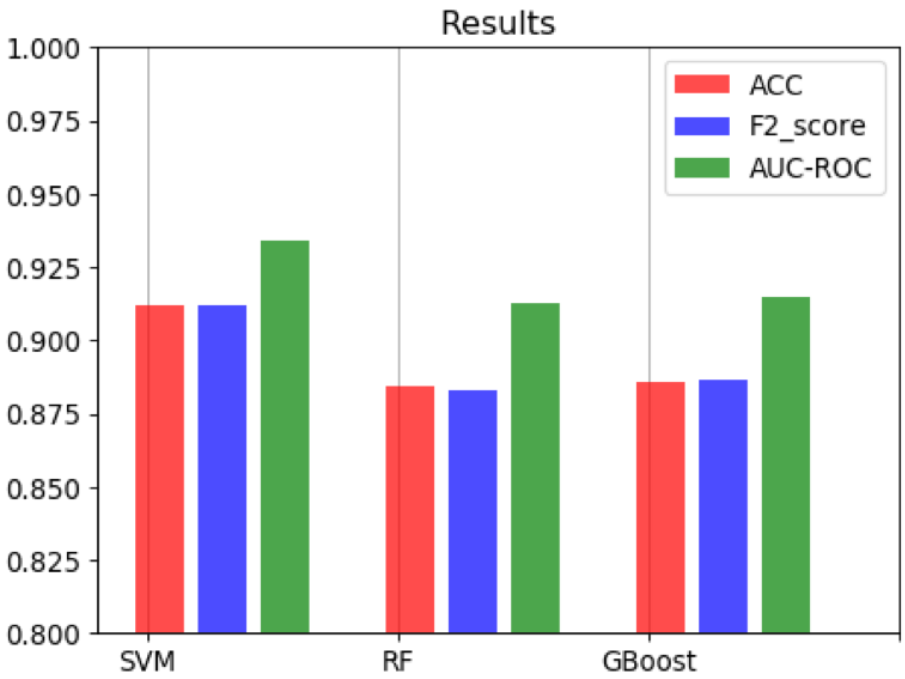


Рис. 1: Results.

5 Набор данных для определения тональности

В данной работе использовался следующий набор данных: отзывы на категории товаров, представленные в разных категориях на сайте Amazon в период с мая 1996 по июль 2014 года.Набор данных был собран в файлы по категориям профессором Стэнфордского университета Джулианом Макаули.

Количество отзывов по классам тональности

Название категории	Положительные	Нейтральные	Отрицательны
"Рецензии на книги"	7 203 909	955 189	738 943
"Кино и Телевидение"	1 289 602	201 302	206 629
"Бытовая техника"	1 356 067	142 257	190 864
"Медицина и здоровье"	279 801	33 254	33 300
"Компакт-диски и винилы"	903 002	101 824	92 766

5.1 Обработка данных

Первым шагом в данном упражнении, переведём оценки из формата 1-5 в бинарный (-1/1), для этого воспользуемся следующей логикой:

- при overall $\in$ (1,2) присвоим значение -1, что будет означать, что отзывы данной категории имеют отрицательную тональность;
- при overall $\in$ (3) не будем присваивать никакого значения, так как мы будем рассматривать задачу бинарной классификации;
- при overall $\in$ (4,5) присвоим значение 1, так как оставляя отзывы с такими оценками, мы можем быть уверены, что пользователь был удовлетворён своей покупкой, а значит отзыв имеет положительную тональность;

Имеем 5 наборов данных, состоящих из отзывов пользователей и их оценок, где в каждом наборе данных содержится приблизительно миллион записей. Так как в задачах машинного обучения важно иметь сбалансированные выборки, а в разных категориях товаров разное количество положительных и отрицательных отзывов, то была написана функция, которая считает в каждом наборе данных количество: отдельно положительных и отдельно отрицательных отзывов, берёт наименьшее из них и обрезает класс, где отзывов больше, тем самым получается равное количество положительных и отрицательных отзывов в выборке.

Далее удаляем из данных знаки препинания и другие лишние символы (отмечу, что они могут быть исследованы при более глубоком анализе) и приводим все слова к нижнему регистру.

6 Анализ тональности

6.1 Векторизация отзывов

Первым этапом при анализе тех или иных текстов методами машинного обучения, текстовые данные требуется перевести в числовой формат.

Метод TF-IDF (TF — term frequency, IDF — inverse document frequency) - метод, суть которого состоит в оценке важности слова как в контексте рассматриваемого экземпляра документа, так и относительно других документов, находящихся в коллекции.

Частота слов (TF) вычисляется как отношение количества вхождений конкретного слова в документе к общему количеству слов в этом документе и записывается в следующем виде:

$tf(a, d) = n_a / \sum_k(n_k)$,

где n_a - количество вхождений слова a в документ n, $\sum_k(n_k)$ - общее число слов в документе n.

Обратная частота документа (IDF) - отношение общего количества документов в коллекции $|D|$ к числу документов, где присутствует данное слово $|d_i \in D|a \in d_i|$ и записывается в виде:

$idf(a, D) = log(|D|/|d_i \in D|a \in d_i|)$

В конечном итоге, меру на которую данный метод опирается, можно записать следующей формулой: $tf - idf(a, d, D) = tf(a, d) * idf(a, D)$

6.2 Логистическая регрессия

Для решения данной задачи классификации на два класса использовалась логистическая регрессия, которая является одним из методов построения линейного классификатора и даёт оценивать апостериорные вероятности принадлежности объектов классам.

Так как логистическая регрессия является частным случаем линейного классификатора, то стоит упомянуть, что главная идея линейного классификатора заключается в том, что признаковое пространство может быть разделено гиперплоскостью на два отдельных полупространства. В каждом из полупространств прогнозируется одно (в случае бинарной классификации, из двух) значение целевого класса. Когда это можно сделать без ошибок, то выборка признаётся линейно разделимой.

6.3 Оценка точности классификации

Исследователями выделяется большое количество оценки точности классификации.

Самым популярным методом является метод Accuracy: суть заключается в том, что при исследовании точности модели мы смотрим на долю правиль-

ных ответов, которая может быть записана в следующем виде: $BaseRate = argmax\frac{1}{l}\sum_{i=1}^{l}[y_o = y_i]$ Данную формулу также принято записывать в следующем виде: $AccuracyRate = \frac{TP+TN}{TP+FP+FN+TN}$

6.4 Результаты

Результаты работы классификатора:

Название категории	AccuracyRate, %
"Рецензии на книги"(рис.2)	91.47%
"Кино и Телевидение"(рис.3)	90.65%
"Бытовая техника"	89.84%
"Медицина и здоровье"(рис.4)	85.78%
"Компакт-диски и винилы"	89.32%

Рис. 2: Метрика ACC для категории: "Рецензии на книги".

Рис. 3: Метрика ACC для категории: "Кино и Телевидение".

Рис. 4: Метрика ACC для категории: "Медицина и здоровье".

7 Визуализация

7.1 Платформа Tableau Public

При решении задачи анализа большого количества данных целесообразно применять удобные средства визуализации. В рамках нашего проекта было решено использовать платформу Tableau Public (public.tableau.com) – бесплатный редактор инфографики, позволяющий представить большинство типов данных в доступной форме. Это современное удобное средство построения разного рода зависимостей с возможностью последующей публикации на веб-ресурсах. Этот сервис позволяет работать с разными форматами данных - Microsoft Exel, Text, JSON, pdf и др. Полученные в результате рисунки автоматички сохраняютя в личном кабинете пользователя, что обеспечивает доступ к ним из любых устройств. Общий объем файлов пользователя ограничивается 10 Гб. Свое изображение можно дополнять текстом, а также прикреплять к нему гиперссылки и другие рисунки (см. рис. 5).

Соотношение занимаемой площади элементов и их порядок задается пользователем. Результат работы можно публиковать на веб-страницах. При этом пользователь может интерактивно взаимодействовать с полотном: выделять зависимости, приближать изображение, изменять интервалы фильтрации.

Приобтаемые навыки работы с платформой при построении графиков ROC-кривых алгоритмов классификации, подбора параметра C и различных вспомогательных изображений будут применяться в визуализации результатов анализа нашего конечного продукта.

7.2 Облака слов

Если алгоритм машинного обучения обрабатывает текстовые данные, полезно иметь представление о них. В частности, для алгоритма логистической регрессии мы строим облака слов [?] – изображение некоторого множества слов в соответствии с их критерием значимости. Критерий значимости определяется алгоритмом машинного обучения и представляет собой дествительное

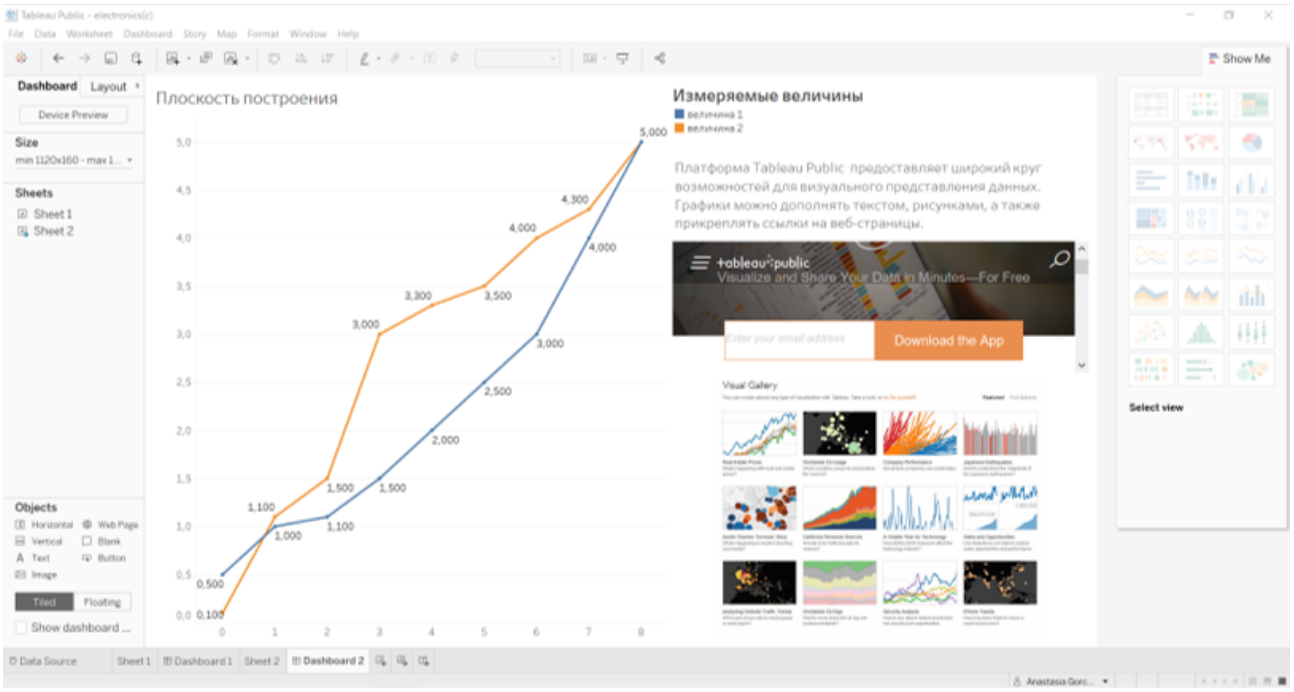


Рис. 5: Типичный вид интерфейса платформы Tableau Public

положительное или отрицательное число.

Построение облака слов в Tableau Public выполнено на примере классификации методом логичтической регрессии. Для каждой категории данных получено изображение, где значимость слова определяется его близостью к центру, размером и интенсивностью оттенка. На рис. приведдены несколько из них для сравнения. В первой паре совпадений крайне мало, хотя обе категории тесно связаны с общим понятием «досуг». Во второй паре совпадений намного больше, хотя категории и кажутся и более отдаленными. Из этого можно сделать вывод, что нельзя выделить какое-то единственное множество слов разумного объема, используемое при описании чего-либо.

7.3 Визуализация процесса обучения

На основе данных, полученных из метода логистическоой регрессии были построены кривые, визуализирующие процесс обучения. На рис. 7.3 видно, как зависит точность классификатора от параметра алгоритма C. Максимальная точность достигается в интервале от 1 до 10.

Для визуального представления точности работы классификатора удобно использовать ROC-кривые. На рис.7.3 такая кривая для категории Health and Personal Care.

Конечно, приведенные зависимости можно построить и пользуясь другими ресурсами, но эта работа была проделана для определения особенностей платформы редактора. На данный момент мы знаем какую структуру должны иметь данные, их объемные ограничения, а также опробовали способы построение наиболее популярных форм визуализации. Полученный опыт бу-

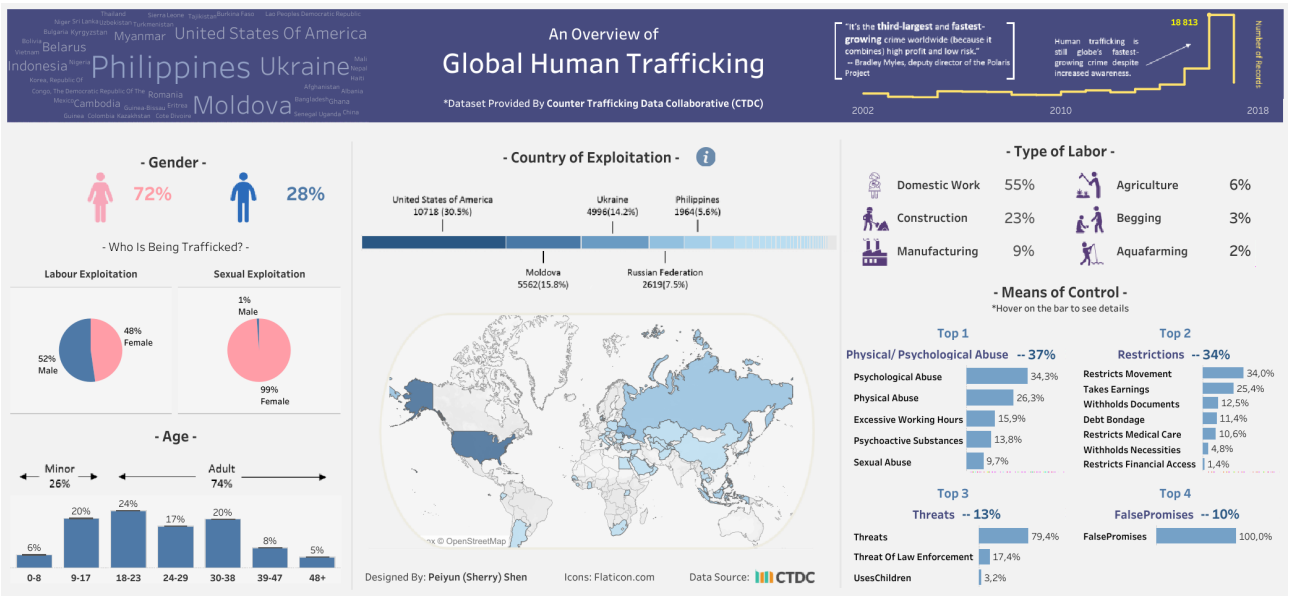
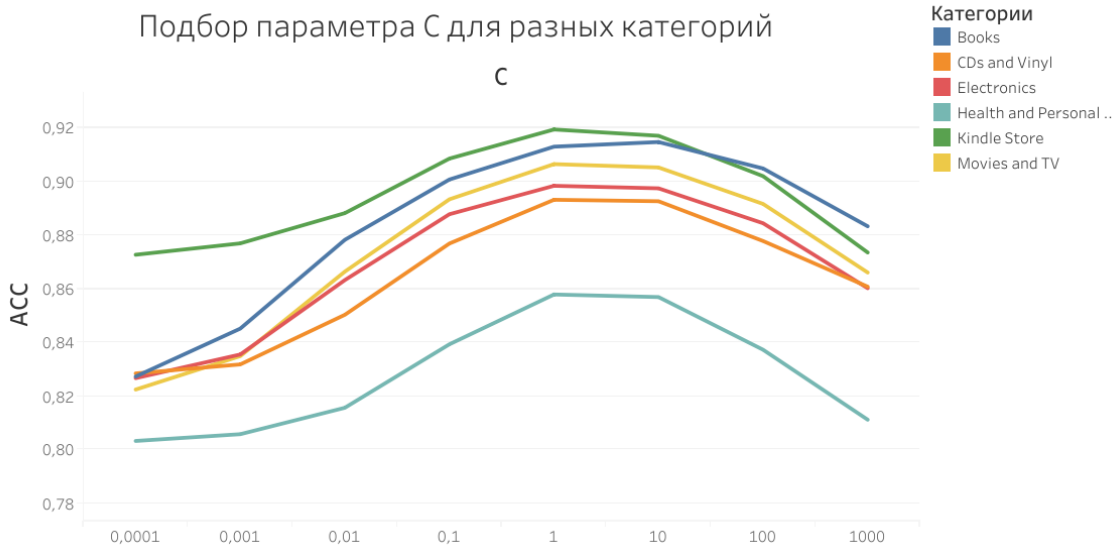


Рис. 6: Пример инфографики, выполненной в Tableau Public

дет применен для оформления конечного результата нашего продукта.

Рис. 7: Визуализация облака слов для разных категорий




ROC-кривая для категории Health and Personal Care
fpr
tpr
1,0
0,8
0,6
0,4
0,2
0,0
0,00
0,01
0,01
0,02
0,02
0,03
0,03
0,04
0,04
0,05
0,06
0,06
0,07
0,08
0,08
0,09
0,10
0,11
0,11
0,12
0,13
0,14
0,16
0,17
0,18
0,19
0,20
0,21
0,22
0,24
0,25
0,27
0,28
0,30
0,33
0,35
0,38
0,41
0,44
0,47
0,52
0,58
0,65
0,80