

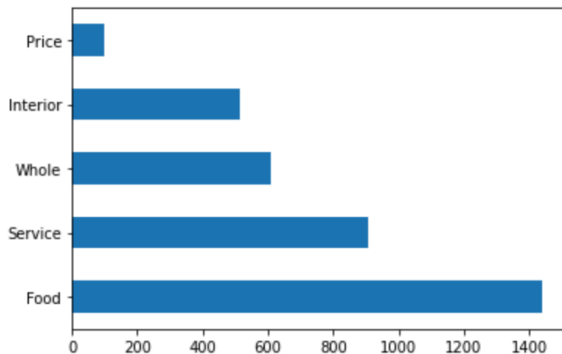
Отчет по проекту

Выполнили Мария Бочарова и Ксения Данилова.

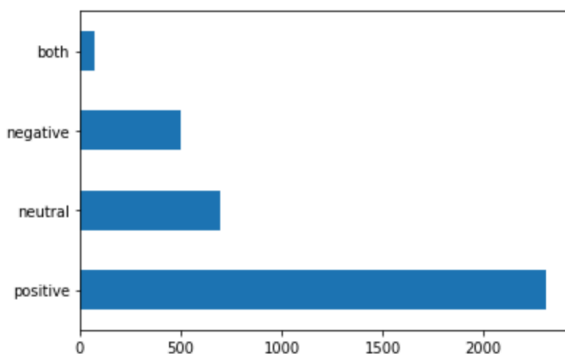
Exploratory Data Analysis

Перед началом работы мы проводим EDA, чтобы лучше узнать датасет, с которым работаем. Результаты анализа представлены в ноутбуке.

Больше всего комментариев в ревью было сделано относительно категории Food. См. график:



Сентименты, встречающиеся в датасете, в основном положительные:



	Baseline	Пополнение словаря по категории Food	Лемматизация текстов ревью	+ Лемматизация словаря	+ SBERT
Full category accuracy	0.4642	0.5004	0.4936	0.4513	0.4936
Partial category accuracy	0.6034	0.6517	0.6534	0.6070	0.6534
Full mention sentiment accuracy	0.6772	0.6223	0.6221	0.5720	0.6156
Partial mention sentiment accuracy	0.5720	0.5545	0.5409	0.5230	0.5308
Overall sentiment accuracy	0.5239	0.5127	0.5155	0.3972	0.5014

Табл. 1

Корпус

В качестве корпуса используется набор отзывов на рестораны. Train корпус состоит из 213 ревью и 3573 выделенных упоминаний по 6 категориям ('Whole', 'Interior', 'Service', 'Food', 'Price') и сентиментам (positive, negative, neutral, both). Бейзлайн выполнен на основе частотного словаря: в него входят только аспекты, встретившиеся в train'e, им приписывается самая частотная категория. Test датасет размечается на основе частотного словаря по полным или частичным совпадениям между ключами словаря и токенами в тексте. Все наши эксперименты выполняются на основе бейзлайна.

Эксперименты

1. Пополнение словаря

В качестве первого эксперимента попробуем расширить словари best_mention_category и best_mention_sentiment, добавив новые наименования блюд и продуктов. Новые наименования категории Food мы парсим с сайта <https://fitaudit.ru/food/abc>, так как готовых датасетов на русском мы не нашли. Всего было извлечено 602 новых наименования, общий объем корпуса составил 1977. Основная проблема в этом подходе — невозможность даже статистически посчитать сентимент. Мы решили поэкспериментировать и для каждого наименования определить сентимент как positive. В итоге, как видно из Табл. 1, этот подход улучшил показатели full и partial category accuracy приблизительно на 5%. В то же время заметно ухудшились метрики full и partial mention sentiment accuracy (на 5,5% и 1,75% соответственно), метрика overall sentiment accuracy снизилась приблизительно на 1%.

2. Использование лемматизации

Лемматизация — один из важнейших этапов препроцессинга. Мы попробовали применить лемматизацию и в данном задании и узнали, как добавление этого шага повлияло на метрики.

Мы добавили лемматизацию на шаге токенизации текста каждого ревью. Мы создаем словарь со следующей структурой: {лемма1: [{токен1: (начало, конец), {токен2: (начало, конец)}}]}. Например, словарь для текста «Рука руку моет» будет иметь следующий вид: {рука: [{рука: (0, 3), {руку: (5, 8)}], моет: [{мыть: (10, 13)}}]. Затем мы искали пересечения ключей данного словаря с best_mention_category и best_mention_sentiment, если есть пересечение, то мы извлекаем из словаря токен и его позицию. Этот шаг, по сравнению с предыдущим экспериментом, не сильно повлиял на значения метрик и даже немного ухудшил их.

Мы также пробовали лемматизировать ключи в best_mention_category и best_mention_sentiment (результаты в колонке «+Лемматизация словаря»), но это существенно понизило ключевые метрики (full и partial category accuracy почти на 5%), поэтому от этого метода мы отказались.

Промежуточный результат

Итак, как мы видим, добавление новых слов в словарь и лемматизация заметно влияют на метрики full и partial category accuracy, увеличивая значения примерно на 5%; в то же время метрики, связанные с sentiment accuracy заметно ухудшились. Попытаемся решить эту проблему.

3. Предобученные эмбединги

Попробуем использовать предобученные эмбединги, чтобы предсказать сентимент для слов из категории Food, которые мы спарсили на кулинарном сайте. До этого мы по

дефолту указывали сентимент «positive», теперь попробуем с помощью предобученных Bert эмбедингов предсказать сентимент. Будем использовать модель модель SBERT (sbert_large_nlu_ru) от Сбера. Схема модели – это классический многослойный перцептрон поверх фичей BERT: mean pooling применяется к последнему encoder-слою, чтобы получить векторы предложений, а сверху добавляются три полносвязных слоя. Результаты тестирования этого метода можно увидеть в Табл. 1 в колонке «+SBERT». Как мы видим, full и partial mention sentiment accuracy при использовании SBERT все равно заметно ниже, чем при использовании бейзлайна. Более того, результаты ниже, чем при использовании дефолтного «positive» для всех новых слов категории Food.

Анализ ошибок

Анализ ошибок проводится на результатах, полученных при выполнении 2-го эксперимента с лемматизацией текстов.

Что касается определения сентиментов, количественно больше всего ошибок было допущено при детекции позитивных комментариев (440 ошибок), а при обнаружении нейтральных, негативных и both — 58, 63 и 24 соответственно. Это распределение вполне соотносится с изначальным распределением размеченных сентиментов: всего в размеченном корпусе содержится 65% позитивных сентиментов.

Ошибки в обнаружении категорий распределились следующим образом: *Food*: 264, *Service*: 158, *Whole*: 101, *Interior*: 53, *Price*: 9.

В train датасете категория Food была представлена 40% примеров, Service — 25%, Whole — 17%, Interior — 14%, а Price — 4%. Распределение ошибок также соотносится с изначальным распределением категорий на тренировочном корпусе.

Что касается лексем, ниже приведены те слова, в определении категорий или сентимент которых чаще всего допускались ошибки:

('меню', 32), ('вкусно', 26), ('человек', 17), ('заказ', 15), ('принесли', 14), ('вкусные', 10), ('впечатление', 9), ('вечер', 9), ('праздник', 9), ('заказали', 8).

Удивительно, что среди лексем фигурирует, казалось бы, такое ярко окрашенное слово, как «вкусно», которое без какого-либо контекста указывает и на категорию (Food), и на сентимент (positive). Единственное, следует также учитывать, что перед «вкусно» стоит отрицательная частица «не», полностью меняющая сентимент.

Итог

Мы смогли улучшить метрики full и partial category accuracy примерно на 5% каждую, используя дополнительные данные из кулинарного сайта, составив 0.5 и 0.65 соответственно. В то же время, ни один контекстно-независимый метод не показал хорошего результата для full, partial и overall mention sentiment accuracy (что грустно, мы очень старались).

Идеи по улучшению

В дальнейшем можно использовать контекстно-зависимые методы для того, чтобы улавливать большее количество информации (как минимум отрицательные частицы). Также могут быть реализованы следующие идеи.

Идея 1: учитывать сентимент предложения, в котором содержится аспектов упоминание.

Идея 2: учитывать сентимент всего текста при определении сентимент упоминания.

Идея 3: использовать сентименты и предложения, и всего ревью для предсказания сентимент упоминания.