Алексеева Анастасия

Астафьев Павел

Дедов Тимофей

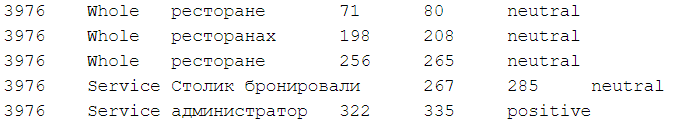
Кузнецова Светлана

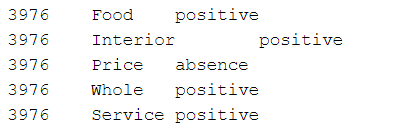
Отчет группы

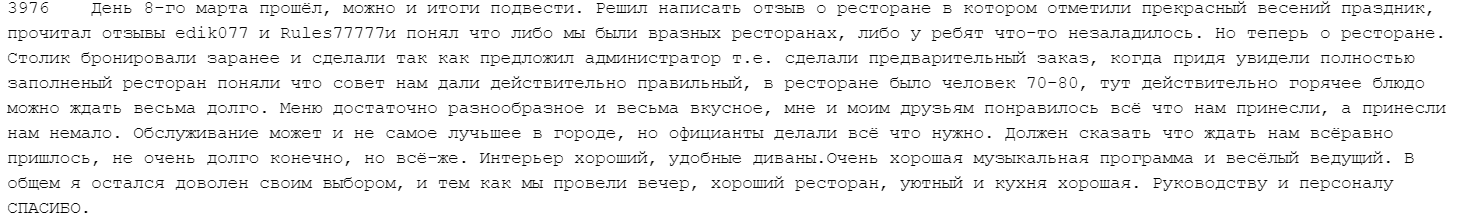
1. Данные

Данные были взяты из предложенных train датасетов, как и разметка.

Для обучения использовались тексты отзывов, в которых уже были выделены важные аспекты, категории аспектов, начало и конец аспекта, а также оценка его тональности.







1. Методы и признаки

Для начальной обработки данных использовалась библиотека stanza для токенизирования и лемматизации.

Далее для каждого предложения мы сделали маски из векторов, где 0 - обычное слово, 1 - слово для разметки категорий.

В качестве модели для выделения сущностей мы выбрали модель Bert для русского языка, потому что она предобучена строить контекстно-зависимые векторные представления токенов на большом корпусе текстов и вполне неплохо с этим справляется, которая затем была дообучена на данных из train датасета.

Для классификации предложений воспользовались векторизацией модели navec, предобученной на русском языке на корпусе художественной литературы. Мы смотрели на Евклидово расстояние между векторами (полученными из navec) выделенных аспектов и слов, выбранных нами как “базовые” для каждой из 5 категорий ('ресторан', 'интерьер', 'еда', 'обслуживание', 'цена' для ‘Whole’, ‘Interior’, ‘Food’, ‘Service’, ‘Price’ соответственно) и определял, к какой категории относится слово.

Для определения тональности мы вновь воспользовались rubert, который возвращал оценку каждого предложения.

1. Результаты тестирования

Мы реализовали решения для всех трех задач разными методами.

Для выделения аспектов мы написали нейросеть на основе берта, для классификации аспектов просто смотрели на вектора, для оценки тональности использовали готовую модель.

Результат выделения аспектов получился не очень хорошим из-за малого количества данных, на которых обучалась модель с бертом. Возможно, в силу малого количества данных, лучше было бы использовать другую модель, например BiLSTM-CRF. Также в качестве усовершенствования проекта можно было бы улучшить функцию классификации, например, не просто смотреть на Евклидово расстояние между векторами двух слов (зачастую не очень показательное), а сделать нейросеть, которая обучится классификации.

1. Квантитативный и качественный анализ ошибок

Мы не использовали предобученные NER модели, так как слишком маленькая выборка, и на ней они бы не дали нормальных результатов, возможно, используя их, мы бы добились лучшего качества (комбинируя 2 способа? или увеличив выборку?). Еще можно было бы использовать алгоритм Reptile, он вроде дает лучшие результаты, чем Bert.