ОТЧЕТ ПО ПРОЕКТУ АВТОБРЕЯ. Выполнено Чижиковой Анастасией и Егоровой Евгенией.

(1) **Корпус и разметка**: мы использовали только предложенные *данные с размеченными аспектами* и ничего не добавляли. Разбиение на train и test было сделано, как в образце
Объем train корпуса: 3573 выделенных аспектов в 213 текстах.
Объем test корпуса: 1190 аспектов в 71 текстах, а также сами тексты отзывов.

Для решения задач использовались следующие готовые модели:

- `"cointegrated/rubert-tiny"` из библиотеки transformers для векторизации нграмм;
- `"blanchefort/rubert-base-cased-sentiment-rusentiment”` из библиотеки transformers для векторизации нграмм для сантимент анализа;

(2) Методология:

Задача 1:

- при помощи nltk-модуля выделяем из тестовых текстов n-граммы;
- векторизуем выделенные и имеющиеся в train датасете аспектные n-граммы;
- попарно находим косинусную близость между выделенными и трэиновым n-граммами;
- оставляем только те выделенные n-граммы из тестовых текстов, для которых в train аспектах нашли очень близкого соседа (“очень близко” определяется установленным пороговым значением косинусной близости);
- ищем отфильтрованные n-граммы в тестовых текстах и приписываем им категорию ближайшего соседа.

Комментарии:

- так как 95% аспектов трэина это уни-, би-, три- и 4-граммы, было решено выделять только 1-, 2-, 3-, 4-граммы из тестовых текстов, предполагая, что в них распределение примерно похожее, то есть мы изначально пренебрегаем примерно 5%;
- были проведены эксперименты с выбором меры ассоциации, количества n-грамм, которые мы выделяем, и трешхолдом, выбраны, на наш взгляд, оптимальные значения;
- перед тем, как искать ближайших соседей из трэин-аспектов, выделенные n-граммы фильтровались по pos-tag схеме: мы оставляли только те n-граммы, pos-tag схема которых встречалась в pos-tag схемах train аспектов (это было

сделано для избавления от нестандартных n-грамм, которые иногда выдавали достаточно высокие значения косинусной близости).

Задача 2:

Провели два эксперимента:

- определение сантимента с помощью предобученной трансформер-модели по ближайшим синтаксическим соседям аспекта
- определение сантимента с помощью предобученной трансформер-модели по всему предложению, в которое входит аспект - второй способ показал лучший результат

- файнтюним готовую модель на наших тренировочных аспектах и определяем сантимент анализ выделенной части - модель показала результаты лучше

Задача 3:

- из модели, использованной для предыдущей задачи, достаем вероятности предсказаний сантимента для каждого;
- формируем предсказание сантимента для категории как взвешенную сумму вероятностей сантиментов аспектов данной категории.

(3) Результаты + анализ ошибок

Задача 1:

По факту мы выделяли все n-граммы из трэина, а также те, которые очень похожи на те, что есть в трэине, но не повторяют их. То есть так мы могли учитывать опечатки или какие-то очень синонимичные слова и добавлять их к трэину, поэтому реколл в сравнении с реколлом бейзлайна мог только увеличиваться, за исключением случаев, когда мы удаляли какие-то n-граммы из-за фильтрации.

Нам удалось поднять реколл, причем таким методом его можно бы было поднять еще больше, однако это бы сильно уменьшало пресижн. Поднять пресижн, нам, к сожалению, не удалось. В некотором роде это можно связать с качеством разметки - иногда некоторые аспекты из того, что является аспектом, не выделяются; некоторые аспекты выделены не полностью (например, из “официант Александра” выделено аспектом “официант”, при этом встречаются и случаи, когда имя тоже входит в аспект). В целом, если смотреть на сами выделенные нами n-граммы, то они выглядят действительно очень похоже на реальные аспекты, и не совсем понятно, почему они не были размечены как таковые.

В итоге были не выделены те аспекты, которые совсем не похожи ни на что. Также некоторой проблемой является невозможность определить правильную ширину разметки (возвращаясь к примеру с официантом) - в случаях, если в n-грамму

входили несколько слов, которые сами по отдельности могут быть аспектом, мы считали аспектом и общую n-грамму, и ее части - это явно влияет на пресижн.

Задача 2:

В такой конфигурации мы объединили с классом neutral те, у которых сантимент both, потому что это самый размытый класс, с которым непонятно как работать.

Чаще всего ошибки в сентименте появляются тогда, когда сентимент выражен неявно и модель с трудом его считывает. Кроме того, хоть способ с выделением полных предложений и показал лучший результат, чем окно из соседей, этот подход не учитывает случаи, когда в предложении есть противопоставление между позитивными и негативными аспектами.

Задача 3:

Изменение мажоритарной системы на взвешенную имеет свои преимущества, однако в качестве весов лучше бы иметь какую-то метрику получше, иначе мы в какой-то степени опираемся на недостатки самой модели. Если модель плохо поняла текст и не очень уверена в предсказаниях, то скорее всего такой подход припишет тексту нейтральный класс, при этом это не всегда будет правдой.