ViaMobi

NLP skills

Dataset

- Отзывы на рестораны
- 47139 строк, 6 столбцов
- 6 классов

	review_id	general	food	interior	service	text
16383	16383	0	5	6	7	Сходили с семьей на ужин . Очередной ФастФуд ...
21095	21095	0	3	6	8	Чтобы не тратить время в " пробке " зашли с сы...
43487	43487	0	0	0	0	Уютно посидеть и поболтать , да ещё и вкусно п...
30038	30038	0	9	8	9	С Антре номер раз у меня отношения сложные . ...
2541	2541	0	6	8	-	Забрела почти случайно , прельстившись " вкусн...

Data preprocessing

1) Пропуски, дубликаты, аномалии
2) Баланс классов
3) Форматирование текста: приведение к нижнему регистру, удаление знаков препинания (исключение “! : - ( )“), исправление грамматических ошибок
4) Удаление стоп-слов (гипотеза проверена-отклонена)
5) Лемматизация

Итог:

- 3199 строк,
- 3 столбца
- 5 классов

general	text	finish_text
5	самый офигенный клуб в москве . таких клубов ...	[самый, офигенный, клуб, в, москва, такой, клу...
5	отмечали юбилей своей мамочки . он был 05.09.2...	[отмечать, юбилей, свой, мамочка, он, быть, пр...
5	решил оставить отзыв , пока впечатления еще св...	[решить, оставить, отзыв, пока, впечатление, е...
5	08.09.12 праздновали в ресторане palati свадьб...	[праздновать, в, ресторан, palati, свадьба, вс...
3	были в кафе пушкин 16 сентября . приводили на...	[быть, в, кафе, пушкин, сентябрь, приводить, н...

Baseline

- Word2Vec
- Расчет среднего вектора отзыва на весах TF-IDF
- Градиентный бустинг над полученными эмбеддингами (n_estimators=100, max_depth=8, learning_rate=1e-1)*
- Разделение на train-test с учетом стратификации таргета (т.к. дисбаланс классов сильный)

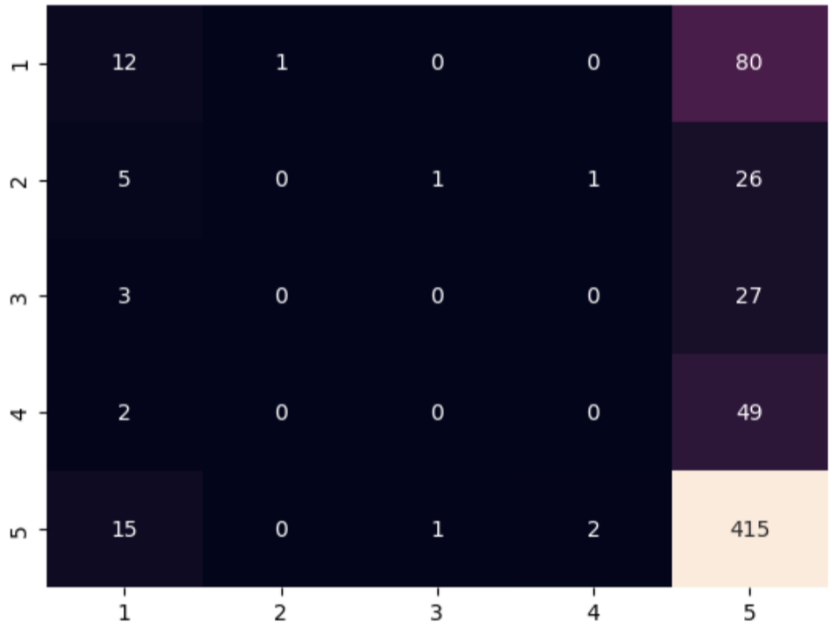
*гиперпараметры не подбирались

Результаты baseline

Метрики качества на тесте

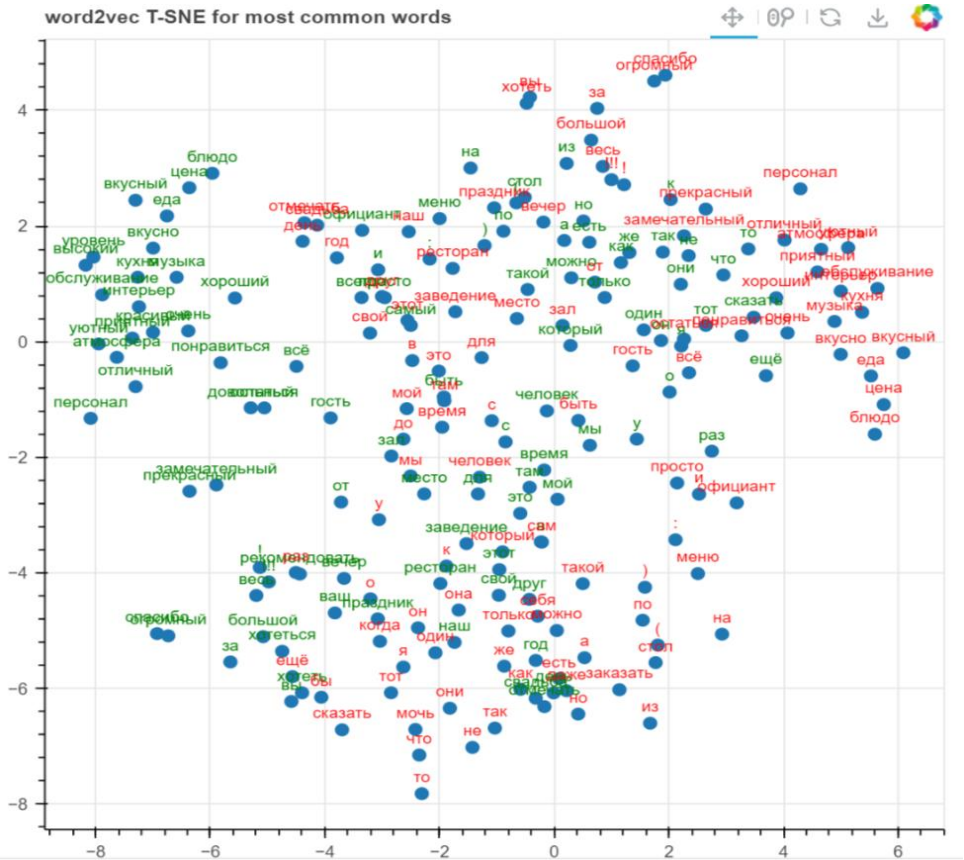
```
Precision:   0.20
Recall:   0.22
F1-measure:   0.20
Accuracy:   0.67
              precision    recall  f1-score   support

           1       0.32      0.13      0.18        93
           2       0.00      0.00      0.00        33
           3       0.00      0.00      0.00        30
           4       0.00      0.00      0.00        51
           5       0.70      0.96      0.81       433

    accuracy                           0.67       640
   macro avg       0.20      0.22      0.20       640
weighted avg       0.52      0.67      0.57       640
```

Confusion Matrix



Визуализация пространства слов (W2V)

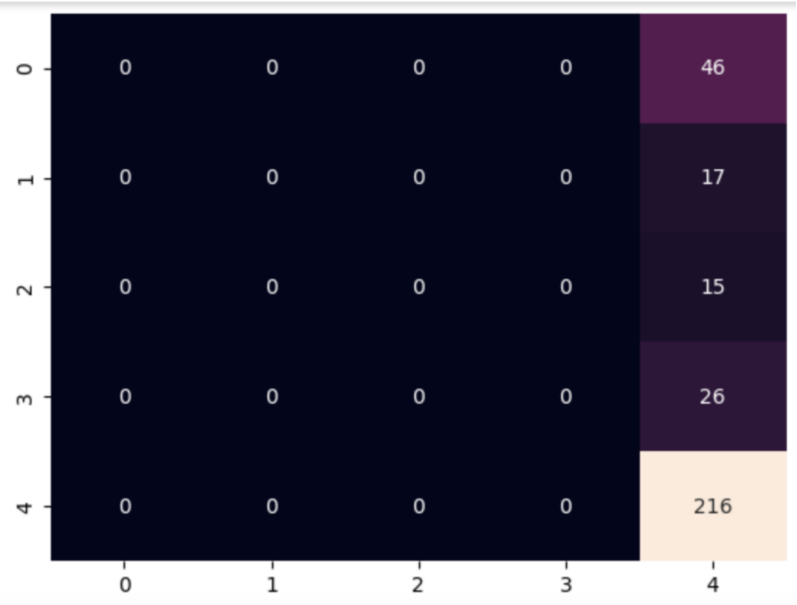
- зеленым - ТОП-100 слов из позитивных отзывов (4 и 5 звезд)
- красным - ТОР-100 слов из негативных отзывов (оценка 3 и ниже)



Fine-tuning ruBERT

Дообучение модели 'ai-forever/ruBert-base' на кастомном датасете под классификацию на 5 классов.

```
Precision:   0.14
Recall:   0.20
F1-measure:   0.16
Accuracy:   0.68
              precision    recall  f1-score   support

           0       0.00      0.00      0.00        46
           1       0.00      0.00      0.00        17
           2       0.00      0.00      0.00        15
           3       0.00      0.00      0.00        26
           4       0.68      1.00      0.81       216

    accuracy                           0.68       320
   macro avg       0.14      0.20      0.16       320
weighted avg       0.46      0.68      0.54       320
```



Выводы

- бустинг над эмбедингами W2V с весами TF-IDF дает более качественный прогноз класса, однако в связи с малым количеством данных не может определить 2 и 3 классы.
- ruBERT не хватило данных для обучения, поэтому в прогнозе только 5 класс. +дообучение было всего 1 эпоха, поэтому результат не удовлетворительный