Определение сантимента компании по новостям

Learn Python 18
Сергей Коньков, Анна Духович
Куратор: Юлдуз Фаттахова

Цель: Создание модели по выявлению общего сантимента компании по новостям

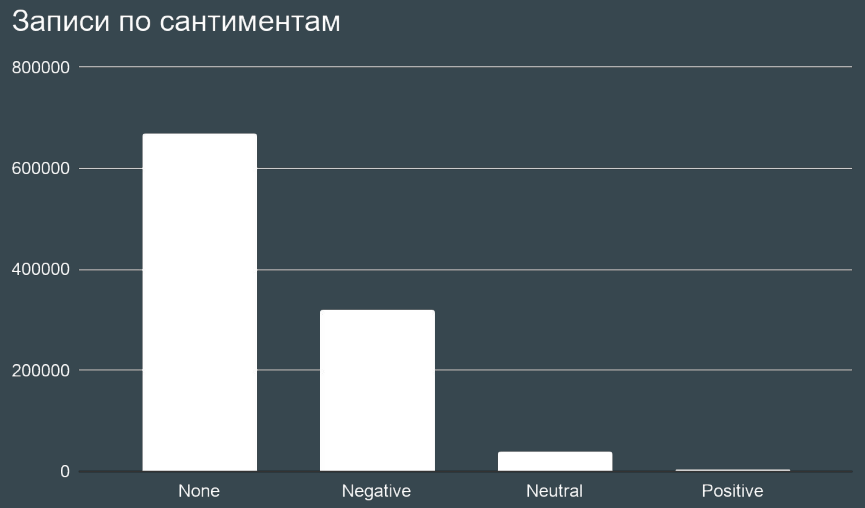
Разделение сантимента на 3 класса:

- Позитивный
- Негативный
- Нейтральный

**сантимент - от фр. ‘sentiment’, буквально переводится, как «чувство» или «настроение».*

Сбор данных

- 300K json документов с kaggle
- Данные размечены некачественно:
  - Сантимент определен некорректно
  - Несуществующие компании
  - Несбалансированные классы
- Разметка данных:
  - Фильтрация текстов по длине
  - Валидация компаний через Wikidata
  - Сравнение API от Google и IBM
  - Разметка с помощью IBM NL
  - Повторная валидация через Wikidata

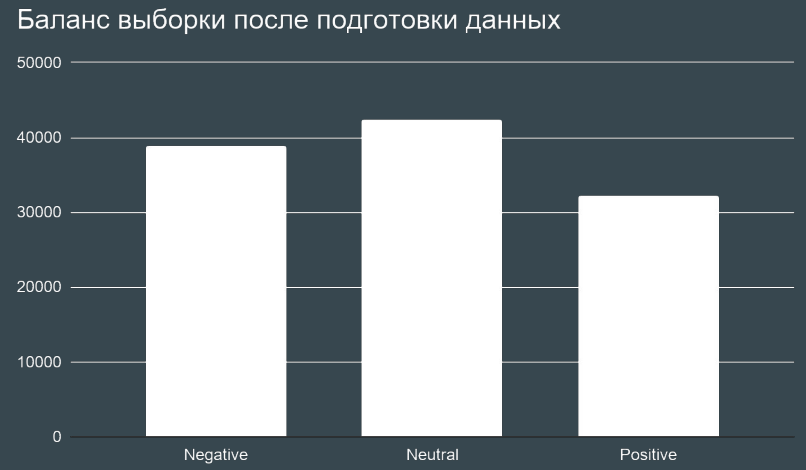


Подготовка данных

1. Очистка текстов
2. Разбиение по предложениям
3. Выделение из текста предложений относящихся к компаниям
4. Проверка сбалансированности выборки

Результат подготовки:

- 113к сегментов текста с упоминанием компаний
- 7.5к уникальных названий компаний
- 40к уникальных документов



Конструирование признаков и моделирование

1

Векторизация текста:

- TF-IDF
- BoW <

2

Древесные алгоритмы:

- Random Forest
- XGBoost
- LightGBM <
- CatBoost

Линейные алгоритмы:

- Логистическая регрессия <
- SVM
- MLP

3

Нетекстовые признаки:

- Random Forest: Feature importance

Метрики:

	precision	recall	f1-score
0	0.85	0.85	0.85
1	0.82	0.88	0.85
2	0.81	0.74	0.77

Прогноз

Сантимент

	neg	neu	pos	All
neg	29%	3%	3%	35%
neu	3%	31%	3%	37%
pos	4%	4%	20%	28%
All	36%	38%	26%	100%

Стек Используемых Технологий

Сбор данных:

- requests
- aiohttp
- json
- csv

Подготовка данных:

- numpy
- pandas
- nltk
- spacy
- re

Конструирование признаков:

- sklearn.decomposition.TruncatedSVD
- sklearn.feature_extraction.text
- sklearn.preprocessing
- scipy

Модели:

- Линейные:
  - sklearn.linear_model.LogisticRegression
  - sklearn.svm
  - sklearn.neural_network.MLPClassifier
- Древесные:
  - sklearn.ensemble.RandomForestClassifier
  - xgboost
  - lightgbm
  - catboost

Визуализация:

- matplotlib
- seaborn

Общего назначения:

- itertools
- os
- sys
- collections
- pickle