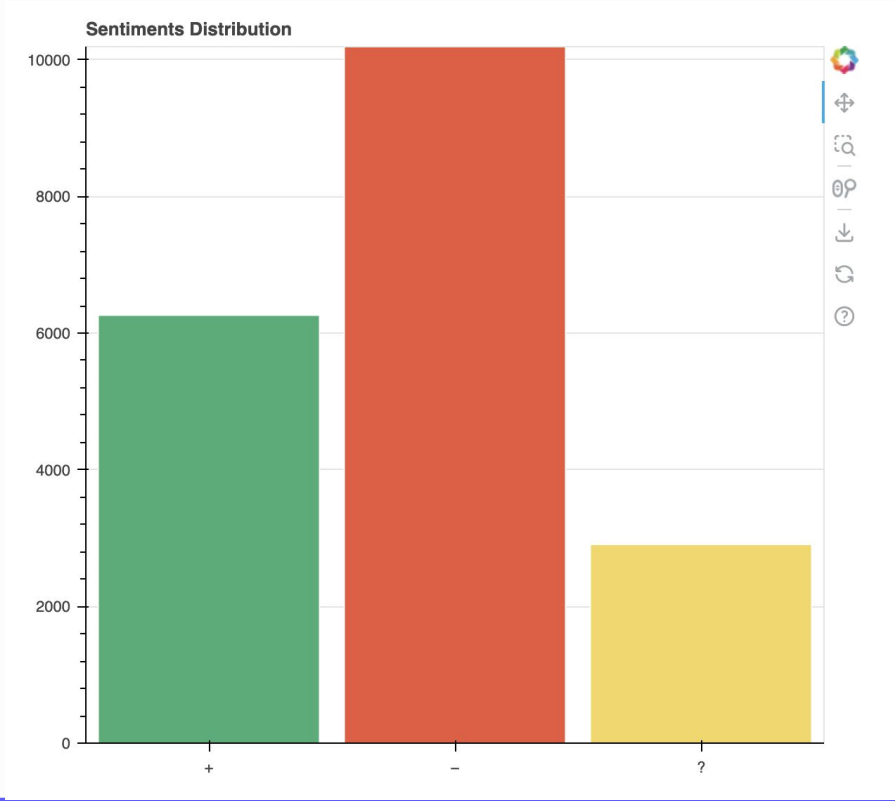
Intro Hook Explore Explain Apply Share Evaluate Expand

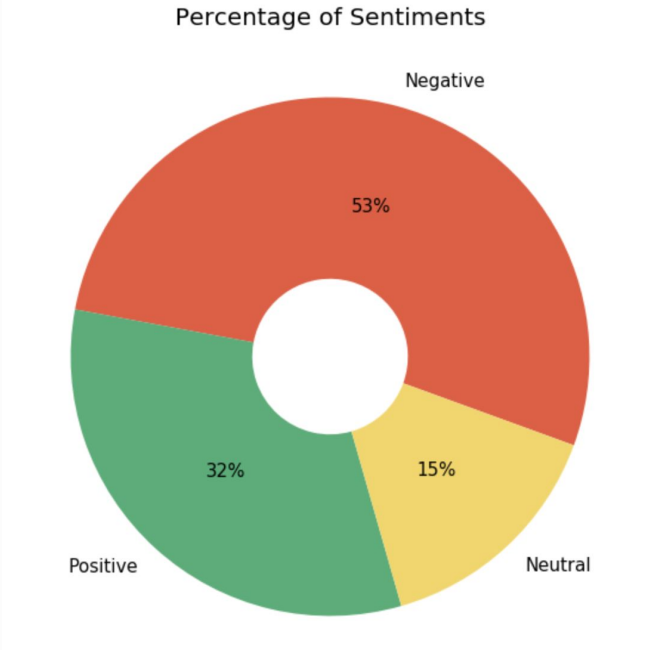
HSE Data Science Hack

Галимзянова Дарья, НИУ ВШЭ
Богданова Анна, НИУ ВШЭ
Лебедева Анна, НИУ ВШЭ

Баланс классов



Баланс классов



Mean word counts

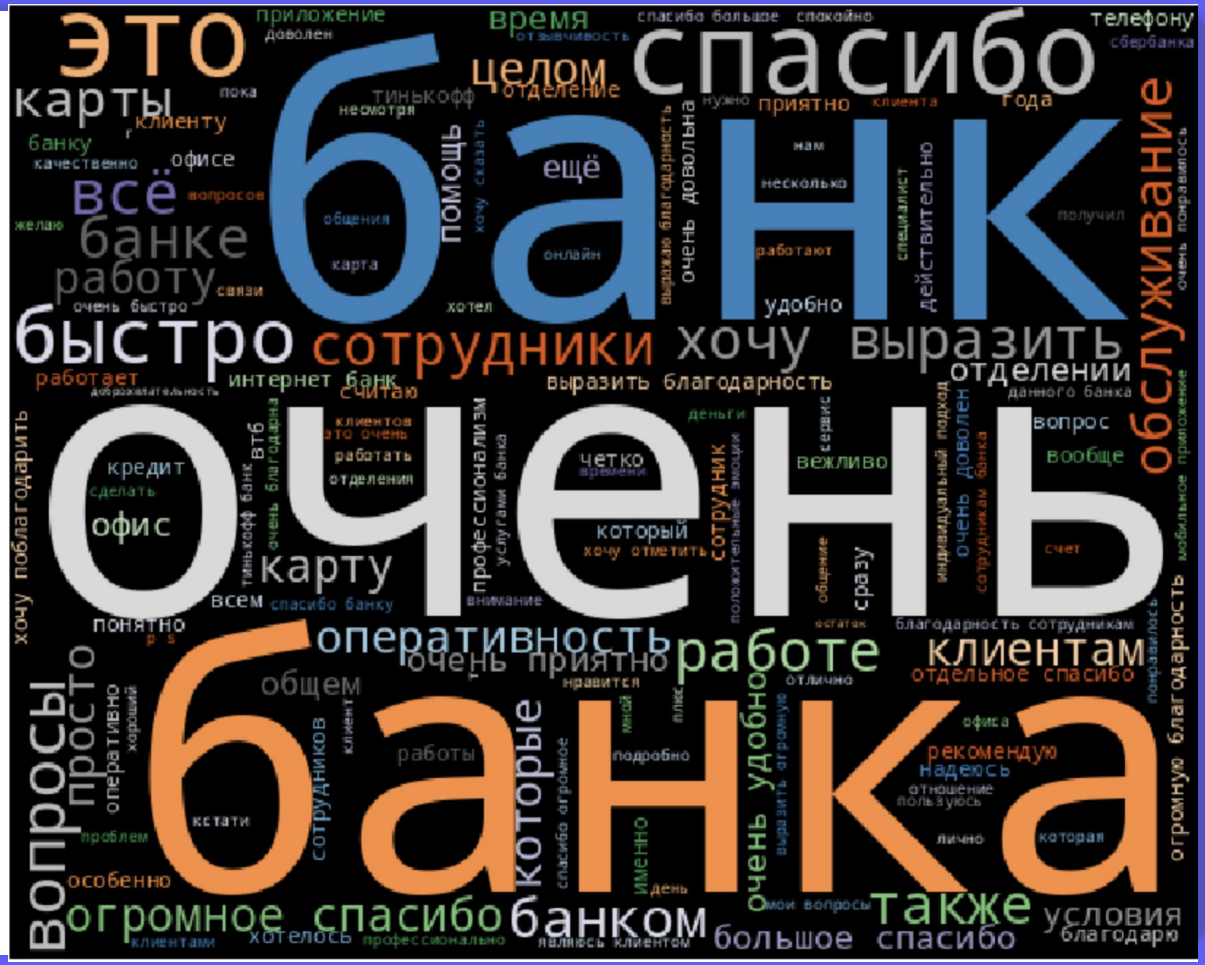
In all categories: **5.97**

Positive: **6.24**

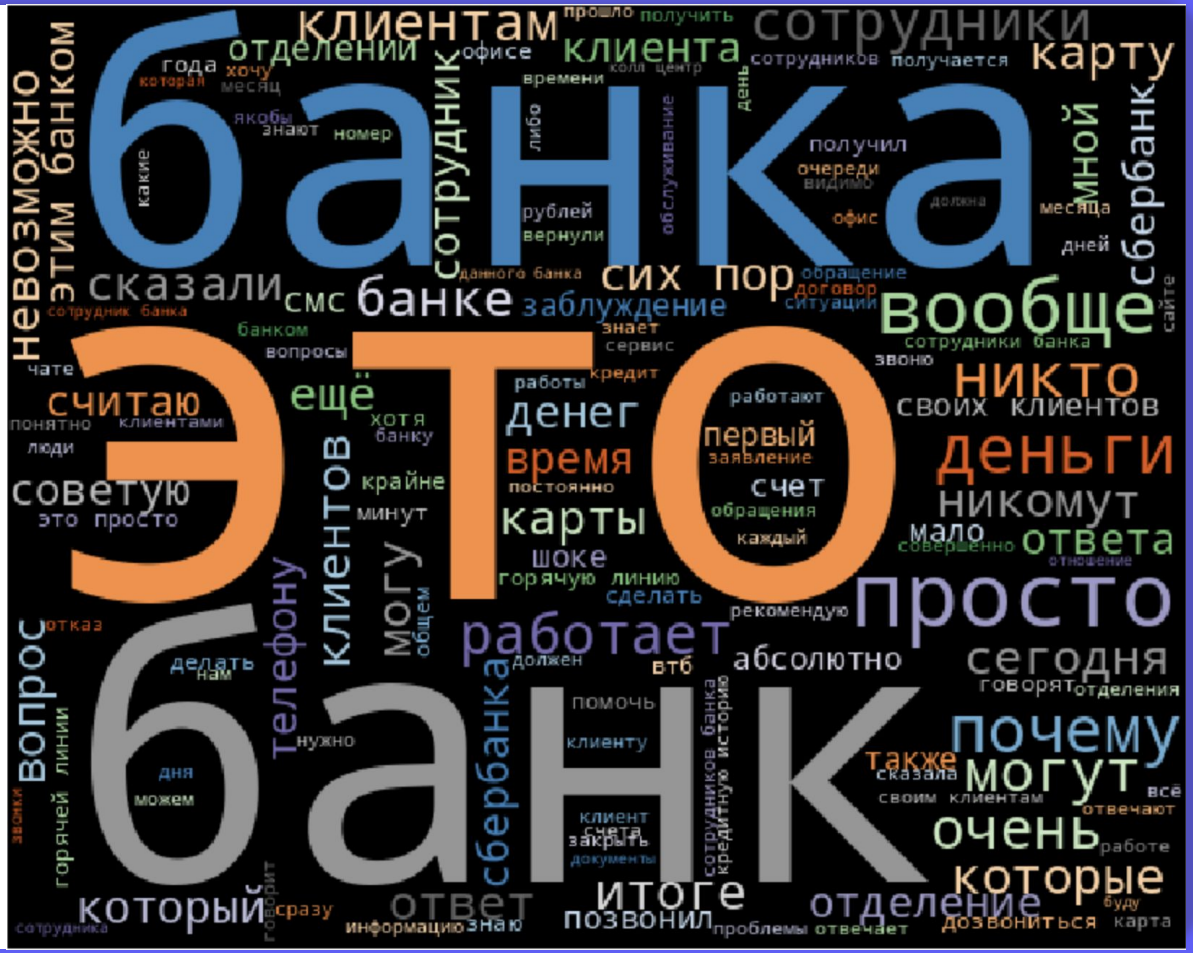
Negative: **5.86**

general word cloud on lemmatized data:

positive messages word cloud:



negative messages word cloud:



Использование Yandex Datasphere

Итоги:

- Создали проект
- Подключили s3 (сохраняли чекпоинты модели)
- Подключили git

Проблемы:

- kernel самопроизвольно перезапускается, значения всех переменных пропадают
- Проблемы с зависимостями

Препроцессинг данных

- Удаление пунктуации
- Лемматизация через pymorphy2
- Векторизация

Бейзлайн

Комбинация	RandomForestClassifier
TfidfVectorizer	0.9289
CountVectorizer	0.9293
TfidfVectorizer w/lemm	0.9314
CountVectorizer w/lemm	0.9332

Сентимент

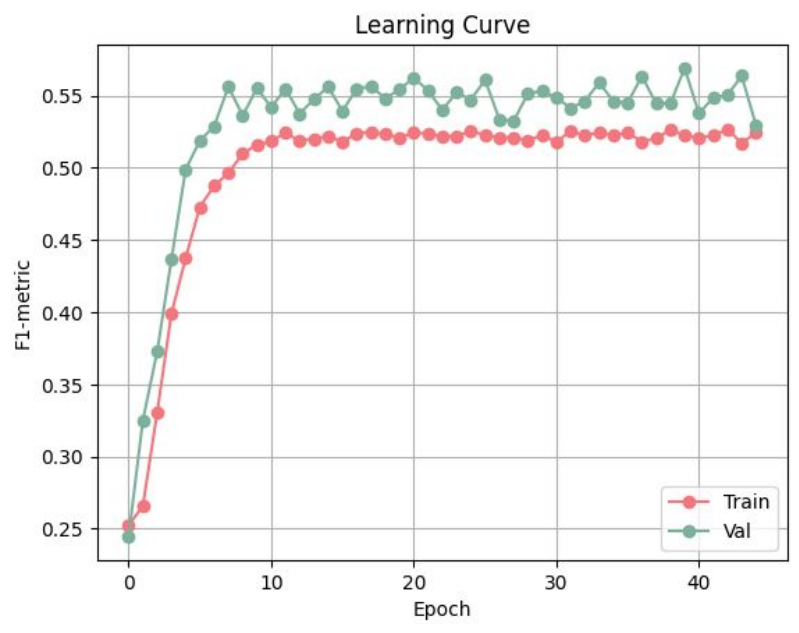
Модель	**ROC-AUC score**
RandomForestClassifier	0.9332
Catboost	0.9304
Catboost + Optuna	0.9279
Rubert finetuned (46 эпох)	~ 0.55

```
0:      learn: 0.6013545        test: 0.5916366 best: 0.5916366 (0)     total: 186ms    remaining: 15m 27s
100:    learn: 0.7813935        test: 0.7392876 best: 0.7392876 (100)   total: 14.6s    remaining: 11m 47s
200:    learn: 0.8153122        test: 0.7640681 best: 0.7656169 (192)   total: 29.2s    remaining: 11m 38s
300:    learn: 0.8395317        test: 0.7862674 best: 0.7862674 (296)   total: 43.6s    remaining: 11m 19s
400:    learn: 0.8570937        test: 0.7924626 best: 0.7950439 (389)   total: 57.4s    remaining: 10m 58s
500:    learn: 0.8702365        test: 0.7976252 best: 0.7976252 (489)   total: 1m 10s   remaining: 10m 35s
600:    learn: 0.8797635        test: 0.8079504 best: 0.8084667 (580)   total: 1m 23s   remaining: 10m 14s
700:    learn: 0.8861915        test: 0.8151781 best: 0.8156944 (691)   total: 1m 37s   remaining: 9m 58s
800:    learn: 0.8921028        test: 0.8187919 best: 0.8187919 (774)   total: 1m 50s   remaining: 9m 40s
900:    learn: 0.8980716        test: 0.8229220 best: 0.8249871 (876)   total: 2m 4s    remaining: 9m 27s
1000:   learn: 0.9018595        test: 0.8327310 best: 0.8332473 (989)   total: 2m 18s   remaining: 9m 12s
1100:   learn: 0.9054752        test: 0.8337636 best: 0.8347961 (1081)  total: 2m 32s   remaining: 8m 58s
1200:   learn: 0.9089761        test: 0.8347961 best: 0.8353123 (1188)  total: 2m 46s   remaining: 8m 45s
1300:   learn: 0.9127640        test: 0.8404750 best: 0.8415075 (1289)  total: 2m 59s   remaining: 8m 31s
1400:   learn: 0.9143136        test: 0.8446051 best: 0.8456376 (1369)  total: 3m 13s   remaining: 8m 17s
1500:   learn: 0.9161501        test: 0.8477026 best: 0.8477026 (1480)  total: 3m 27s   remaining: 8m 3s
1600:   learn: 0.9174702        test: 0.8487352 best: 0.8487352 (1596)  total: 3m 41s   remaining: 7m 50s
1700:   learn: 0.9189624        test: 0.8502839 best: 0.8502839 (1662)  total: 3m 56s   remaining: 7m 38s
1800:   learn: 0.9204545        test: 0.8492514 best: 0.8502839 (1662)  total: 4m 11s   remaining: 7m 26s
1900:   learn: 0.9218893        test: 0.8513165 best: 0.8518327 (1856)  total: 4m 26s   remaining: 7m 14s
2000:   learn: 0.9228076        test: 0.8523490 best: 0.8528653 (1963)  total: 4m 41s   remaining: 7m 2s
2100:   learn: 0.9231520        test: 0.8513165 best: 0.8528653 (1963)  total: 4m 56s   remaining: 6m 49s
2200:   learn: 0.9243572        test: 0.8523490 best: 0.8528653 (1963)  total: 5m 11s   remaining: 6m 36s
2300:   learn: 0.9251033        test: 0.8549303 best: 0.8554466 (2282)  total: 5m 27s   remaining: 6m 23s
2400:   learn: 0.9258494        test: 0.8569954 best: 0.8575116 (2397)  total: 5m 42s   remaining: 6m 10s
2500:   learn: 0.9261364        test: 0.8554466 best: 0.8580279 (2433)  total: 5m 57s   remaining: 5m 57s
2600:   learn: 0.9263659        test: 0.8559628 best: 0.8580279 (2433)  total: 6m 11s   remaining: 5m 42s
2700:   learn: 0.9268251        test: 0.8554466 best: 0.8580279 (2433)  total: 6m 26s   remaining: 5m 28s
2800:   learn: 0.9272268        test: 0.8554466 best: 0.8580279 (2433)  total: 6m 41s   remaining: 5m 14s
2900:   learn: 0.9272842        test: 0.8554466 best: 0.8580279 (2433)  total: 6m 56s   remaining: 5m 1s
Stopped by overfitting detector  (500 iterations wait)

bestTest = 0.8580278782
bestIteration = 2433

Shrink model to first 2434 iterations.
```

	precision	recall	f1-score	support
0	0.87	0.90	0.89	626
1	0.65	0.49	0.56	291
2	0.89	0.94	0.91	1020
accuracy			0.86	1937
macro avg	0.81	0.78	0.79	1937
weighted avg	0.85	0.86	0.85	1937

RuBERT fine-tuning

- DeepPavlov/rubert-base-cased
- Замораживание весов предобученной модели, дообучение классифицирующей головы (два линейных слоя), использование StepLR
- Количество эпох: 45



Категории

- Preprocessing:
  - токенизация,
  - лемматизация,
  - удаление стоп-слов,
  - TfIdfVectorizer,
  - MultiLabelBinarizer

- RandomForestClassifier(n_estimators=300)

ROC-AUC: 0.825