



ІІТМО

**Определение катастрофизации с
использованием технологий больших
данных**

Леонов В.В., Базарова В.В., J4111



обработка и разметка исходных текстовых данных
+
использование моделей классификации для определения
катастрофизации

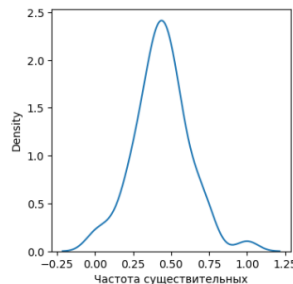
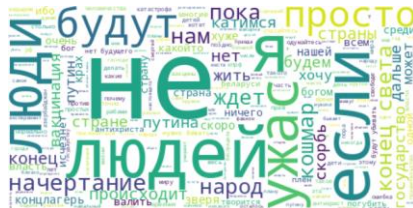
Пример: всё ужасно, нам всем конец!

Обработка и разметка данных

YouTube (@tvrain, 1.5 млн комментариев)

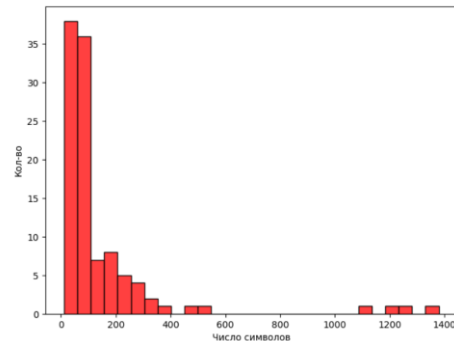
разметка dev-сета вручную

- токенизация, лемматизация, удаление стоп-слов
- определение наиболее частых униграмм, биграмм
- распределения числа символов, частей речи
- определение sentiments и наиболее частых пар эмоций



```
sentiment
neutral      53
negative     46
positive      5
skip          2
speech        1
```

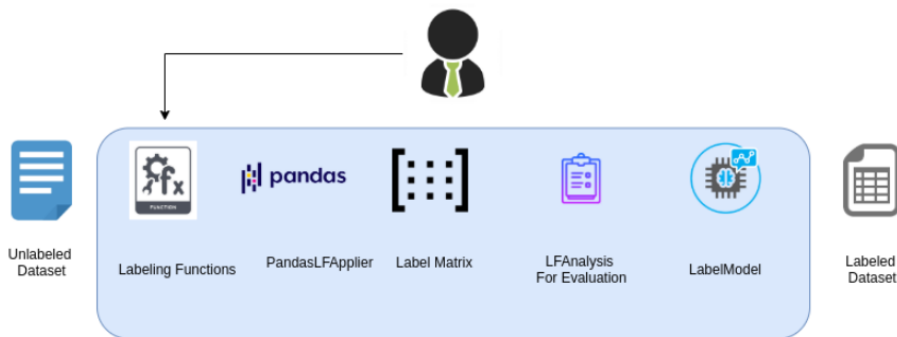
Гистограмма распределения кол-ва букв в комментариях с катастрофизацией (gt)



```
emotion
no_emotion anger      27
anger sadness          9
no_emotion fear        8
no_emotion sadness     7
sadness anger          7
```

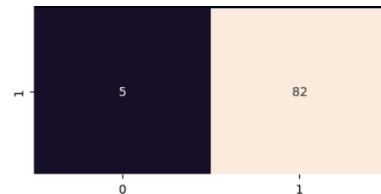
Обработка и разметка данных

Snorkel



```
@labeling_function()
def lf_keywords_catast(x):
    keywords_catast = ['катастрофа', 'ужас', 'ужасно', 'конец', 'проблема',
                       'паника', 'истерика', 'невозможно', 'гибель', 'трагедия',
                       'кончено', 'страх', 'умереть', 'геноцид', 'смерть', 'бояться',
                       'пережить', 'страшно', 'пропасть', 'крах', 'необратимый',
                       'губить', 'концлагерь', 'шок', 'коммар', 'истребить',
                       'катастрофический', 'фатальный', 'непоправимый', 'ужасающий',
                       'хаос', 'хаотичный', 'путеший', 'безнадёжный', 'коммарный']
    return CATASTROPHIZING if any(token in x.normalized_tokens for token in keywords_catast) else ABSTAIN
```

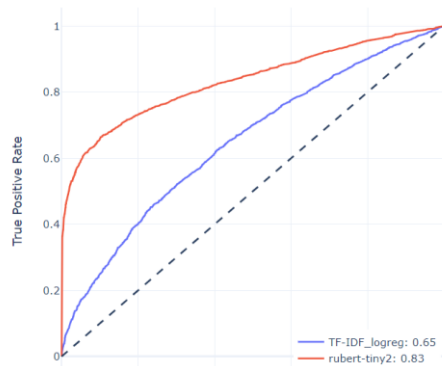
	j	Polarity	Coverage	Overlaps	Conflicts	Correct	Incorrect	Emp. Acc.
lf_keywords_catast	0	[1]	0.560748	0.271028	0.112150	60	0	1.0
lf_regex_catast	1	[1]	0.420561	0.252336	0.093458	45	0	1.0
lf_emotion_not_catast	2	[0]	0.196262	0.149533	0.149533	0	21	0.0



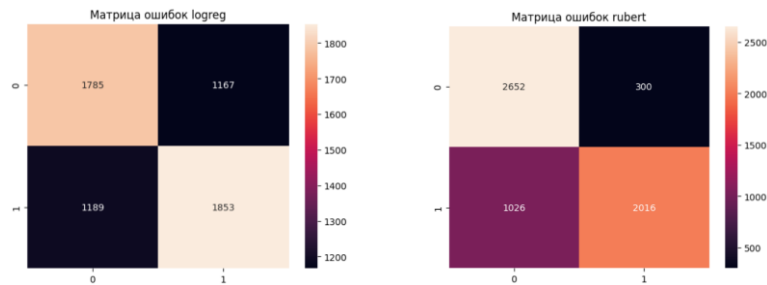
Сравнение моделей

- логистическая регрессия
- дообученный rubert-tiny2

AUC-ROC



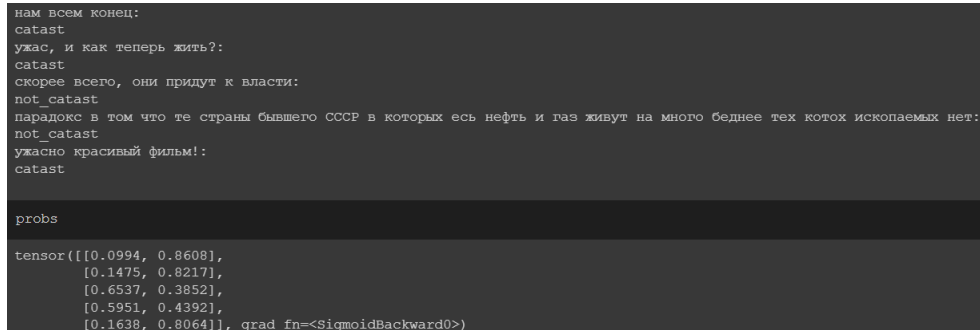
Test:



Dev:



Модель	precision	recall	f1-score
LogReg	0.61	0.61	0.61
rubert-tiny2_catast	0.87	0.66	0.75



```
нам всем конец:
catast
ужас, и как теперь жить?:
catast
скорее всего, они придут к власти:
not_catast
парадокс в том что те страны бывшего СССР в которых есь нефть и газ живут на много беднее тех котох ископаемых нет:
not_catast
ужасно красивый фильм!:
catast

probs

tensor([[[0.0994, 0.8608],
          [0.1475, 0.8217],
          [0.6537, 0.3852],
          [0.5951, 0.4392],
          [0.1638, 0.8064]], grad_fn=<SigmoidBackward0>])
```

Недостатки:

- неточность авто-разметки (правила)
- сравнить больше моделей

Спасибо
за внимание!

it's **MO** *re than a*
UNIVERSITY