# Анализ тональности текста с помощью модели LSTM

Забродина Татьяна МСМТ-213

# Обзор данных

| | Социальная сеть | ID поста | ID комментария | Владелец | ID Владельца | Текст | Эмоциональный окрас | Дата | Лайков |
|---|---|---|---|---|---|---|---|---|---|
| **0** | Вконтакте | -115807015_2334 | 5412482_2336 | Валентин Точилкин | 5412482.0 | Задумка понравилась! Думаю будет смотреться лу... | Нейтральность | 2021-12-31 08:19:05 | 1 |
| **1** | Вконтакте | -30666517_1774956 | 308730199_1774960 | Денис Гончаров | 308730199.0 | Вам тоже здоровья и удачи в прошлом и крепкого... | Вежливость | 2021-12-31 07:09:40 | 9 |
| **2** | Вконтакте | -30666517_1774956 | 10784303_1774979 | Олег Кирин | 10784303.0 | Полагаю: Вера и Надежда – это тимлид и прожект... | Юмор | 2021-12-31 09:20:39 | 4 |
| **3** | Вконтакте | -30666517_1774956 | 586953361_1775092 | Замбек Замбеков | 586953361.0 | этот интеллект на али давно существует | Нейтральность | 2021-12-31 20:38:00 | 1 |
| **4** | Вконтакте | -30666517_1774956 | 19496621_1775350 | Александр Бирабиджанов | 19496621.0 | Будет вечный Свет. | Нейтральность | 2022-01-03 00:04:52 | 1 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **21493** | Вконтакте | -24682865_28855 | 608612_29743 | Алексей Рысаков | 608612.0 | С мая месяца мурыжат с возвратом денег. Все за... | Негатив | 2021-12-08 19:23:12 | 1 |
| **21494** | Вконтакте | -24682865_28855 | 514006271_30353 | Red Kirill | 514006271.0 | Нет войне! | Нейтральность | 2022-03-01 14:40:12 | 0 |
| **21495** | Вконтакте | -139121250_17896 | 551359045_17980 | Владимир Прин | 551359045.0 | Чот много смартфонов на винде я погляжу | Нейтральность | 2021-09-05 19:02:41 | 0 |
| **21496** | Вконтакте | -139121250_17896 | 510436278_17982 | Дима Асеев | 510436278.0 | На тот момент система была хорошей, и мне каже... | Нейтральность | 2021-09-05 19:06:47 | 8 |
| **21497** | Вконтакте | -718901_3582 | 558639831_3603 | Алексей Павловец | 558639831.0 | Мой сын ему 12 лет самостоятельно изучает JS. ... | Нейтральность | 2021-09-10 20:28:53 | 0 |

21498 rows × 9 columns

Для проведения анализа взяты колонки Текст и Эмоциональный окрас

# Проблемы данных

| Эмоциональный окрас | Текст |
|---|---|
| Вежливость | 652 |
| Негатив | 3950 |
| Нейтральность | 11811 |
| Неопределенность | 1146 |
| Позитив | 2539 |
| Юмор | 1399 |

мало

много

Дисбаланс

Плюс: не ясно, что за класс «Неопределенность» -> на время обучения избавимся от него

# Проблемы данных

| index | Текст | Эмоциональный окрас |
|---|---|---|
| 52 | Чаще всего загадывают бабу или мужика. Инфа 100% | Юмор |
| 53 | еще одна кодогенерация😅 убейте | Негатив |
| 54 | Ростелеком, сразу минус! | Нейтральность |
| 55 | > За ваш успех отвечает Ростелеком. для ответа дождитесь освободившего оператора | Юмор |
| 56 | Звучит как угроза. | Нейтральность |
| 57 | Хех. Стажировка - строить великий Российский файерволл а-ля Золотой щит? | Юмор |
| 58 | Даже отписаться захотелось после такого вброса спама =( | Негатив |
| 59 | Заслуженно. | Позитив |
| 60 | С наступающим Новым годом!🎊🎄🥰 Вы большие молодцы - так держать👍💪✌ | Вежливость |

| | | |
|---|---|---|
| Склейки текста | –> | необходимо добавить пробелы |
| Много смайлов | –> | удалить редкие смайлы |
| Много знаков препинания | –> | удалить пунктуацию |
| Разные регистры | –> | привести все слова к нижнему регистру |
| Встречаются одни и те же слова через «е» и «ё» | –> | заменить все на «е» |
| В тексте комментариев есть ссылки | –> | заменяем их на «URL» |
| Много стоп-слов | –> | чистим от них текст |
| Разные формы слов | –> | удаляем окончания (стеммизация) |

# Пример предобработки текста

'Как строки с числами работают🤣🤣'

привели к нижнему регистру
и устранили склейку

'как строки с числами работают 🤣 🤣'

удалили стоп-слова

'строки числами работают 🤣 🤣'

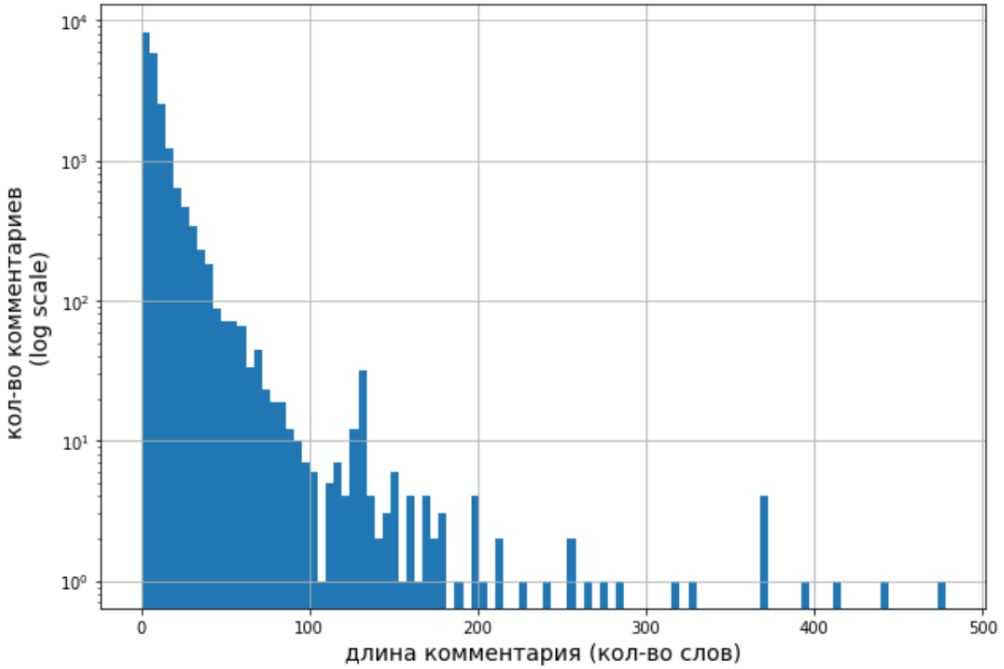
стеммизация

'строк числ работа 🤣 🤣'

Итого:
уникальных слов, смайлов, эмодзи: 23046

# Кодировка текста

выбираем максимальную длину фраз = 100 слов




добавляем паддинги из нулей

```
[[    0     0     0 ...   811   172 10016]
 [    0     0     0 ...   391   856   158]
 [    0     0     0 ...     1  1990 10017]
 ...
 [    0     0     0 ...   814   439 21590]
 [    0     0     0 ...   190    84   300]
 [    0     0     0 ...  1612   721    45]]
```

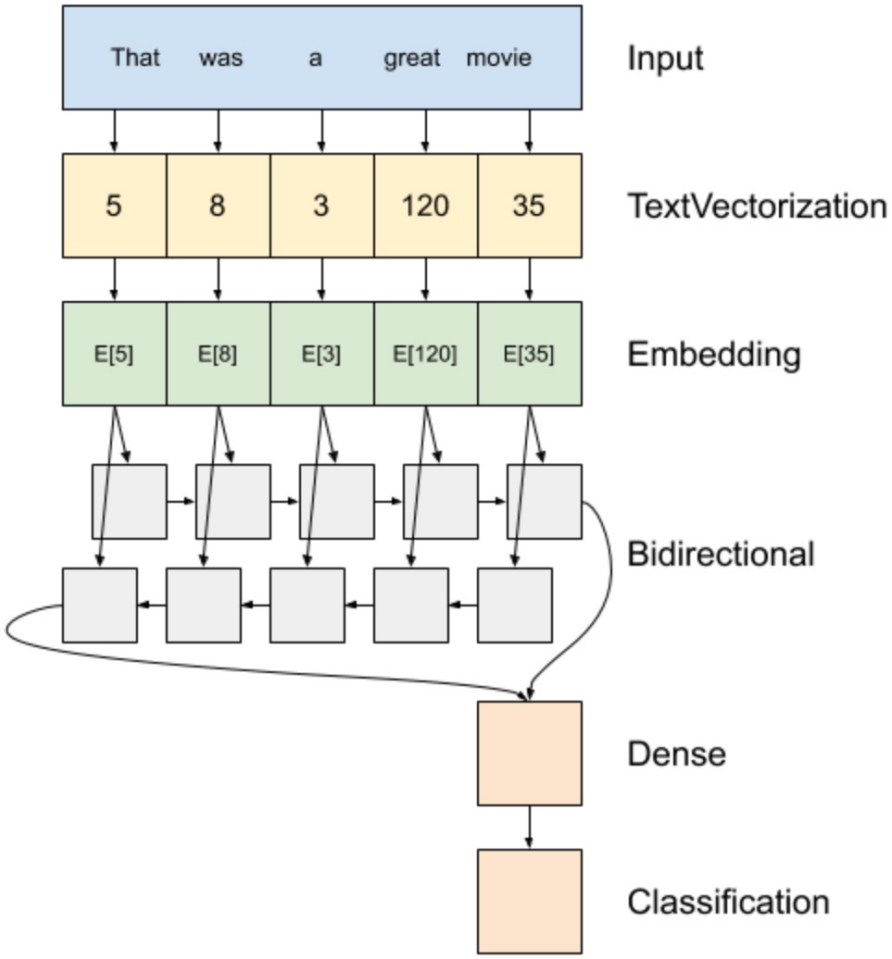
Пример

'строк числ работа 🤣 🤣'

[[812, 243, 20, 71, 71]]

# Модель

```
model = tf.keras.Sequential()

model.add(tf.keras.layers.Embedding(maxWordsCount, 128,
input_length = max_text_len))

model.add(tf.keras.layers.Bidirectional(tf.keras.layers.LSTM(64,
activation='tanh',  return_sequences=True)))

model.add(tf.keras.layers.Bidirectional(tf.keras.layers.LSTM(32,
activation='tanh')))

model.add(tf.keras.layers.Dense(5, activation='softmax'))

model.summary()
```



# Dostoevsky

```
from dostoevsky.tokenization import RegexTokenizer
from dostoevsky.models import FastTextSocialNetworkModel

tokenizer = RegexTokenizer()
FastTextSocialNetworkModel.MODEL_PATH = 'fasttext-social-network-model.bin'
model = FastTextSocialNetworkModel(tokenizer=tokenizer)
```

```
1 model.predict(['Как строки с числами работают🤣🤣'])
```

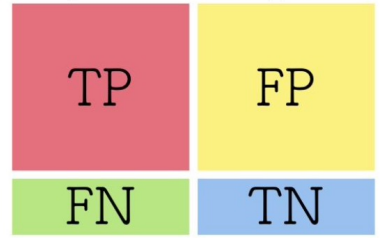
```
[{'negative': 0.08036746829748154,
  'neutral': 0.9511522054672241,
  'positive': 0.0031826822087168694,
  'skip': 0.06755668669939041,
  'speech': 0.0059210699982941 15}]
```

# Результаты

| | accuracy score | f1 score | precision score | recall score |
|---|---|---|---|---|
| **LSTM** | 0.4671 | 0.4544 | 0.5344 | 0.4671 |
| **dostoevsky with preprocess** | 0.4855 | 0.3255 | 0.4333 | 0.4855 |
| **dostoevsky** | 0.4823 | 0.3189 | 0.3057 | 0.4823 |

High precision, low recall


Low precision, high recall



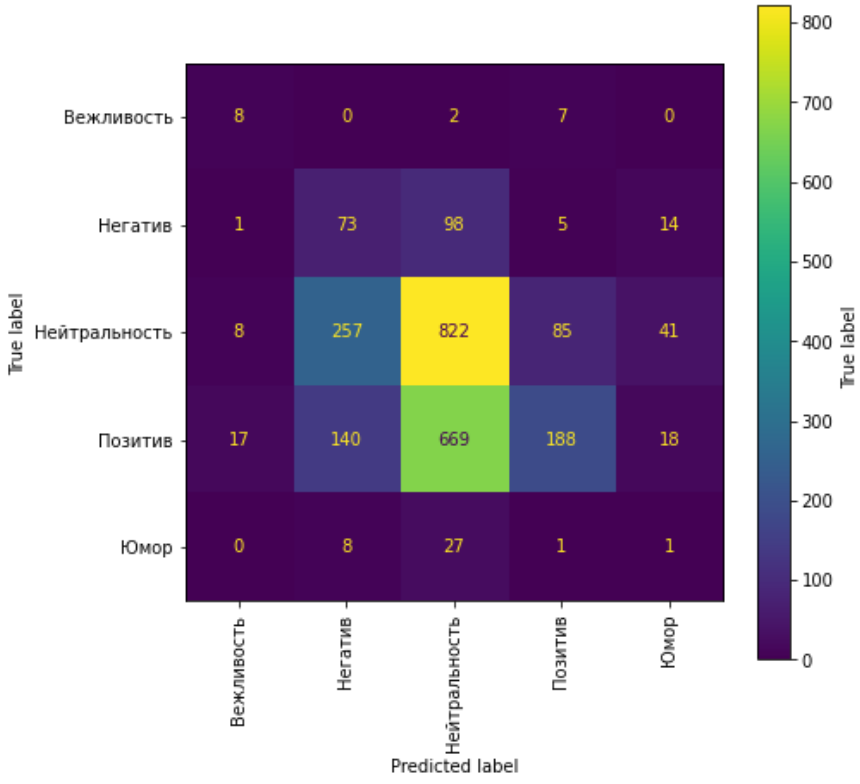

$$\text{Precision} = \frac{TP}{TP + FP} \qquad \text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1 Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

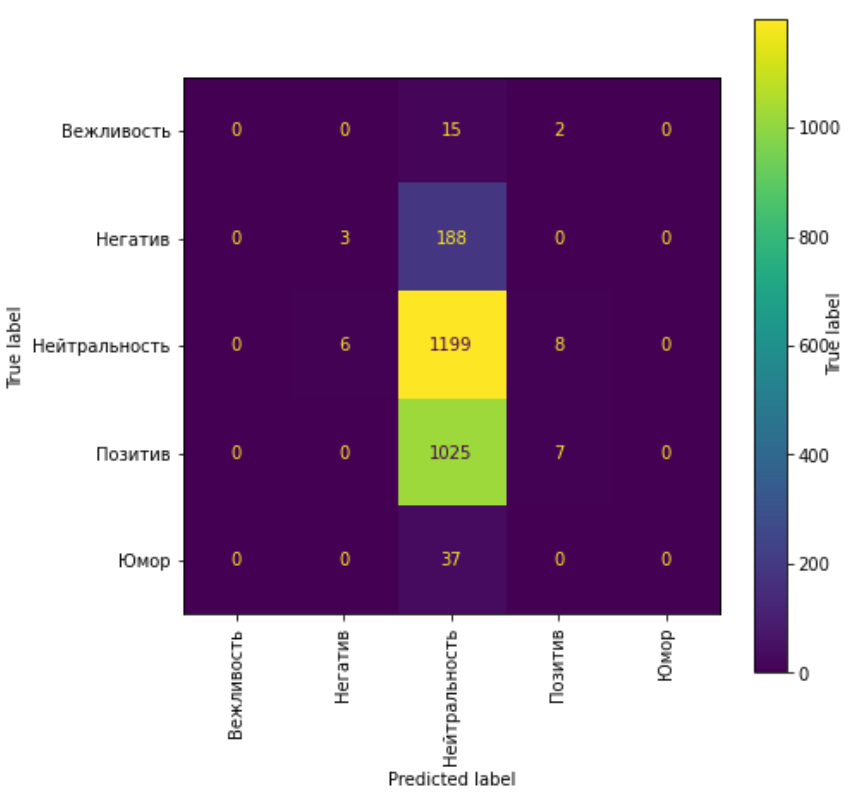
# Результаты

| | Вежливость % | Негатив % | Нейтральность % | Позитив % | Юмор % |
|---|---|---|---|---|---|
| **LSTM** | 47 | 38 | 67 | 18 | 2 |
| **dostoevsky with preprocess** | 0 | 1 | 98 | 0 | 0 |
| **dostoevsky** | 0 | 0 | 98 | 0 | 0 |

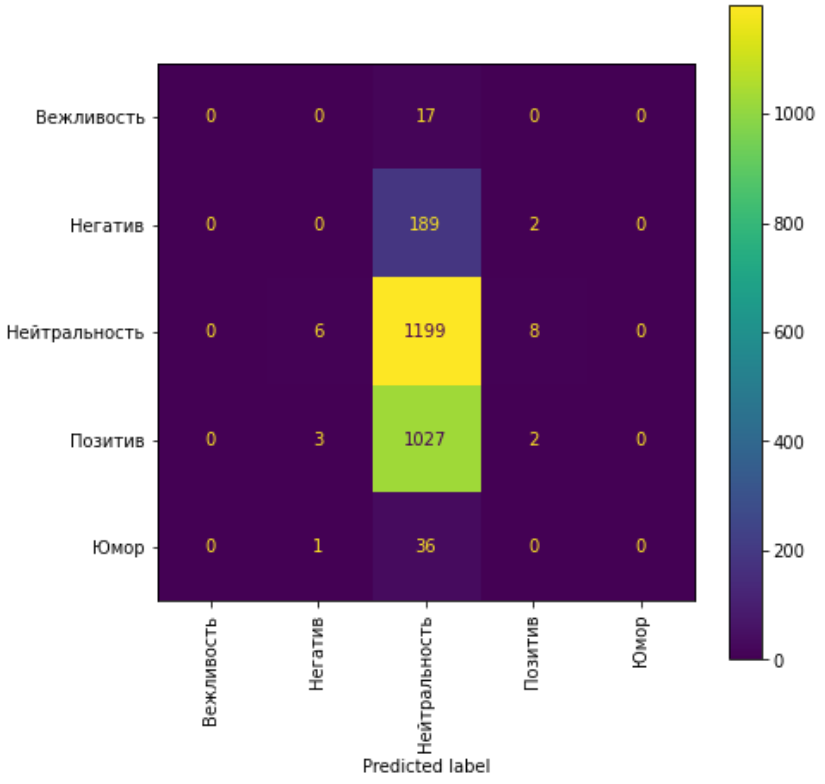
## LSTM




## Dostoevsky (preprocess)




## Dostoevsky



Спасибо за внимание!