



HIGHER SCHOOL
OF ECONOMICS

Regretful Parents

Анализ Reddit постов родителей, которые столкнулись со сложностями и сожалениями после заведения детей.

Алина Тимашова

Компьютерная лингвистика

Содержание

- Описание проекта
- Цели
- Задачи
- Гипотезы
- Методология
- Сложности
- Первые итоги
- Маркеры родителей
- Статистический анализ
- Результаты



Содержание

Описание

Анализ 1007 постов на Реддите из ветки
r/regretfulparents с целью выяснить – о чем
жалеют родители.

Цель проекта

Определить, что беспокоит родителей и есть ли отличия между сожалениями матерей и отцов.

Глобальная цель – исследовать эмоциональную составляющую заведения детей как еще одну причину демографического кризиса.

Содержание

Задачи

- Собрать корпус постов с помощью библиотеки PRAW.
- Составить список маркеров и определить возможный пол автора поста.
- Проанализировать частотные языковые паттерны и сравнить их по родителям.

Гипотезы

[Содержание](#)



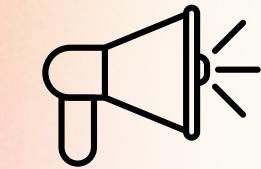
Матери сожалеют чаще

Постов от матерей
будет больше, чем
от отцов.



Сожаления матерей отличаются

Мы будем наблюдать
пересечение, но
не совпадение.



Эмоции тоже отличаются

Возможно, что
у матерей эмоции
будут в спектре грусти,
а у отцов — злости.

Методология

[Содержание](#)

PRAW

- Специальная библиотека для парсинга материалов с Реддита

Re

- Регулярные выражения для маркеров родителей

Pandas

- Данные и таблицы

NLTK, SpaCy

- Предобработка текстов

Sklearn

- Частотный анализ

TextBlob, BERT

- Сентимент-анализ и тематический анализ

Сложности и ограничения

[Содержание](#)

Ограничение количества постов

- Библиотека PRAW позволяет парсить только до 1000 единиц контента. К сожалению, библиотека Pushshift, которая позволяла эти ограничения обойти, теперь доступна только модераторам Реддита.

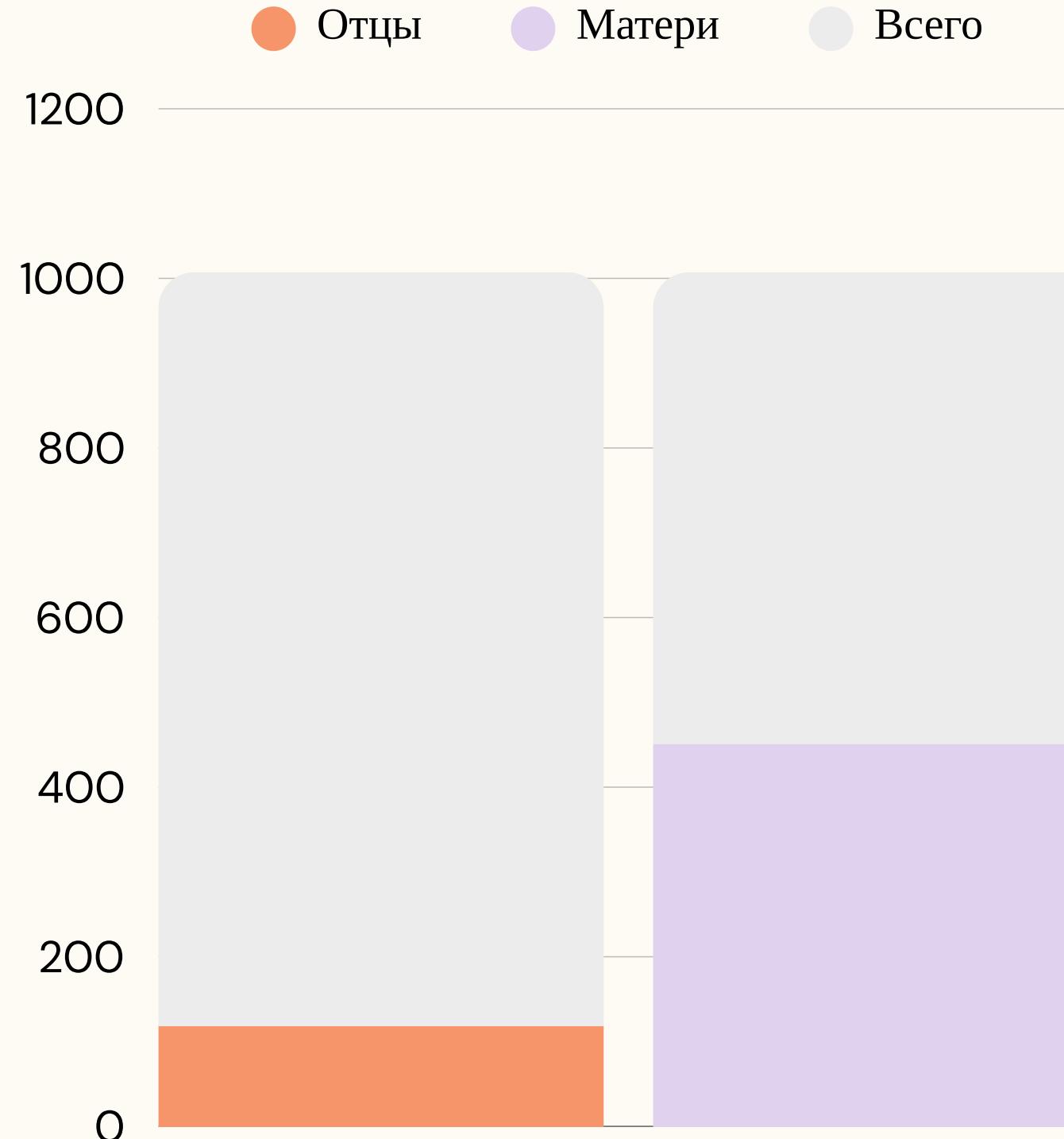
Угадывание родителя

- Реддит ограничивает доступ к информации о пользователях, поэтому мы определяем родителя с определенной долей вероятности.

Неконсистентный язык

- Так как корпус собран на базе мировой соцсети, людям свойственно ошибаться, использовать сокращения или слэнг. Это может затруднить статистический анализ.

Содержание



Первые итоги

Матери почти в 4 раза чаще сожалеют о родительстве.

- Из 1007 постов удалось идентифицировать 570 текстов, 451 из которых – матери, 119 – отцы.

Здесь можно увидеть как рос список маркеров. Благодаря этому число идентифицированных текстов выросло с 406 до 570.

```

1 import re
2
3 mother_markers = [
4     r'I(\\'m|\sam)\s*(\w+\s)*?[Mm](uo)m|other',
5     r'\b([Tt]heir|[Hh]is|[Hh]er)\s*(\w+\s)*?([Dd]ad|[Ff]ather)',
6     r'F\s*\d+\b',
7     r'[Mm](uo)m|other)\sto\b',
8     r'([Mm]y)\s*\.*?\s*([Hh]usband|[Bb]oyfriend|[Ff]iance|[Bb]f)',
9     r'I\s*(\\'m|am|was|got)\s*pregnant',
10    r'[Mm]y\s*pregnancy',
11    r'I\s*gave\s*birth',
12    r'I\s*delivered\s*(a|the)?\s*baby',
13    r'[Aa]s\s*a\s*woman'
14 ]
15
16 father_markers = [
17     r'I\s*(\\'m|am)\s*(a|the)?\s*([Dd]ad|[Ff]ather)',
18     r'\b([Tt]heir|[Hh]is|[Hh]er)\s*([Mm]om|[Mm]other)',
19     r'M\s*\d+\b',
20     r'([Mm]y)?\s*\w+?\s*\b([Ww]ife|[Gg]irlfri',
21     r'[Mm]y\s*(wife|girlfriend)\'??\s*pregnancy',
22     r'[Mm]y\s*(wife|girlfriend|fiance)\s*('s|is|was|got)\s*pregna
23     r'[Aa]s\s*a\s*man',
24 ]

```

```

1     mother_markers = [
2         r'\bi\s*(\w+\s)*?([\\'\']?m|\sam|was|have\sbeen|have\sto\be)\s*(a|the)?\s*
3         (\w+\s)*?(m[uo]m|mother|stahm)\b',
4         r'becoming\s*(a|the)\s*(m[uo]m|mother|stahm)',
5         r'i\s*(\w+\s)*?([\\'\']?m|\sam|was|have\sbeen)\s*(\w+\s)*?
6         \d+\s*months\spostpartum',
7         r'\b(their|his|her|child[\\'\']?s|baby[\\'\']s|son[\\'\']?s|daughter[\\'\']?s)\s*
8         (\w+\s)*?(dad|father)',
9         r'((\bf\s?[\/\ ]?\s?\d+\b)|(\b\d+\s?[\/\ ]?\s?f\b))',
10        r'i?\s*(has|got)?\s*(ppd|postpartum\sdepression)',
11        r'mother\sto\b',
12        r'to\s*be\s*(a|the)?\s*(mother|m[ou]m)',
13        r'my\s*\.*?\s*(husband|boyfriend|fiance|bf)',
14        r'i\s*(\w+\s)*?([\\'\']?m|am|was|got)\s*pregnant',
15        r'my\s*pregnancy',
16        r'i\s*gave\s*birth',
17        r'i\s*delivered\s*(a|the)?\s*baby',
18        r'(as|being)\s*(a|the)?\s*(woman|m[uo]m|mother)',
19        r'\bi\s*.*?had\s*(a|my)?\s*(\w+\s)*?(son|daughter|baby)',
20        r'i\s*(became|have\sbecome)\s*(a|the)?\s*m(other|om(my))',
21        r'i\s*hate\s*motherhood',
22        r'i\s*hate\s*being\s*(a|the)?\s*(m[uo]m|mother|stahm|wife)',
23        r'baby\s*daddy',
24        r'(my)?\s*(he|husband|boyfriend|fiance|bf)\s*(\w+\s)*?want(s|ed)\s*(kids?
|child(ren)?)'
25    ]
26
27    father_markers = [
28        r'i\s*([\\'\']?m|am)\s*(a|the)?\s*(\w+\s)*?(dad|father|stahd)',
29        r'becoming\s*(a|the)\s*(dad|father|stahd)',
30        r'\b(their|his|her|child[\\'\']?s|baby[\\'\']s|son[\\'\']?s|daughter[\\'\']?s)\s*
(m[uo]m|mother)',
31        r'((\bm\s?[\/\ ]?\s?\d+\b)|(\b\d+\s?[\/\ ]?\s?m\b))',
32        r'father\s*to\b',
33        r'my\s*(\w+\s)*?\s*\b(wife|girlfriend|gf)',
34        r'(my)?\s*(wife|girlfriend|gf|partner)[\\'\']?s\s*(\w+\s)*?
(pregnancy|ppd|postpartum\sdepression)',
35        r'(my)?\s*(she|wife|girlfriend|fiance|gf|partner)\s*([\\'\']?
s|is|was|got)\s*preg(nant)?',
36        r'(my)?\s*(she|wife|girlfriend|fiance|gf)\s*(has|got)?\s*(ppd|postpartum
depression)',
37        r'(my)?\s*(she|wife|girlfriend|fiance|gf)\s*(\w+\s)*?want(s|ed)\s*(kids?
|child(ren)?)',
38        r'(as|being)\s*(a|the)\s*(man|dad|father)',
39        r'i\s*hate\s*being\s*(a|the)?\s*(father|dad|husband)',
40        r'i\s*hate\s*fatherhood',
41        r'to\s*be\s*(a|the)?\s*(father|dad)'
42    ]

```

Статистический анализ

[Содержание](#)

```
first_try  
  
def negative_words(text):  
    sentences = re.split(r'[\n\.!\?]', str(text))  
    negative_words = []  
  
    for sentence in sentences:  
        sentiment = TextBlob(sentence).sentiment.polarity  
        if sentiment < 0:  
            negative_words.extend(clean_words(sentence))  
    return negative_words
```

Первая попытка

Сентимент-анализ +
простой частотник.

```
second_try  
  
def apply_tfidf_filtering(all_posts, top_n_words=100):  
    texts = []  
    for post in all_posts:  
        words = clean_words(post)  
        if words:  
            texts.append(' '.join(words))  
  
    vectorizer = TfidfVectorizer(max_features=top_n_words, ngram_range=(1, 2),  
stop_words=stop_words)  
    vectorizer.fit_transform(texts)  
    return set(vectorizer.get_feature_names_out())
```

Вторая попытка

Мешок слов + TF-IDF.

```
third_try  
  
representation_model = KeyBERTInspired()  
topic_model_mother = BERTopic(language="english",  
nr_topics=nr_topics, representation_model=representation_model)  
topic_model_father = BERTopic(language="english",  
nr_topics=nr_topics, representation_model=representation_model)
```

Третья попытка (успешная)

Тематический анализ
BERT.

Результат TF-IDF

[Содержание](#)

Матери

TIME

LIFE

HUSBAND

SON

FAMILY

WORK

DAUGHTER

HOME

JOB

SCHOOL

Отцы

TIME

LIFE

WIFE

WORK

SON

FAMILY

HOME

DAUGHTER

SCHOOL

NIGHT

Результат BERT

Матери

- **Topic -1:** daycare, parenting, husband, life, toddler, constantly, support, family, work, break
- **Topic 0:** parenting, support, depression, health, pregnant, pregnancy, family, care, child, parent
- **Topic 1:** parenthood, birth, stress, pregnant, depressed, pregnancy, child, feel, felt, cry
- **Topic 2:** bedtime, toddler, nap, parenting, bed, husband, baby, alarm, tired, sleeping
- **Topic 3:** daycare, health, overwhelmed, family, living, life, stressed, selfish, miserable, 5yo
- **Topic 4:** hate, mother, mom, life, dread, toddler, pregnant, health, baby, caring
- **Topic 5:** adhd, tantrum, child, kid, autism, mom, whining, autistic, swear, dad

Отцы

- **Topic -1:** parenting, life, stress, support, parent, wife, daughter, family, child, father
- **Topic 0:** misery, life, miserable, resentment, selfish, living, child, live, kid, hate
- **Topic 1:** parenting, care, life, child, therapy, kid, unhappy, nap, break, family
- **Topic 2:** depression, fatherhood, parenting, depressed, pregnant, selfish, life, pregnancy, abortion, suicide

Что дальше?

[Содержание](#)

- Растить корпус вручную
- Анализ семантических полей
- Анализ по возрасту детей