СОЗДАНИЕ НОВОСТНОГО AI-БОТА

или как избавиться от информационного шума

Дата 24.09.2023

КОМАНДА M&L

Науменко
Дмитрий,
ML

Кутькина
Татьяна,
ML

Дамдинов
Зорикто,
Капитан, ML

ПРОБЛЕМАТИКА

1. Слишком много источников ненужной читателю информации

2. Слишком много дубликатов новостей

3. Думскроллинг

4. Отсутствие единого новостного агрегата, способного персонализировать новостные предпочтения читателя

ЗАДАЧИ

Необходимо разработать 2 алгоритма:

1. Алгоритм, способный классифицировать новости по категориям
2. Алгоритм, способный точно и быстро идентифицировать дубликаты новостей

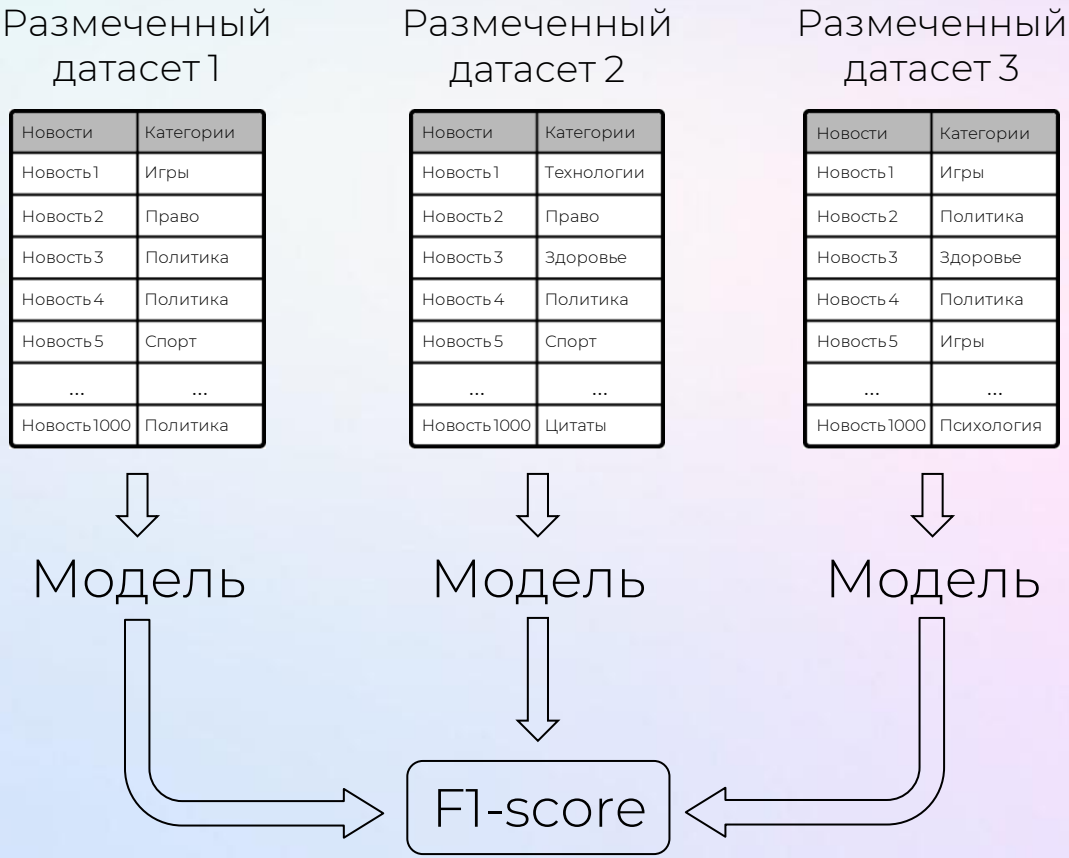
ДАННЫЕ

Способы улучшения датасета:

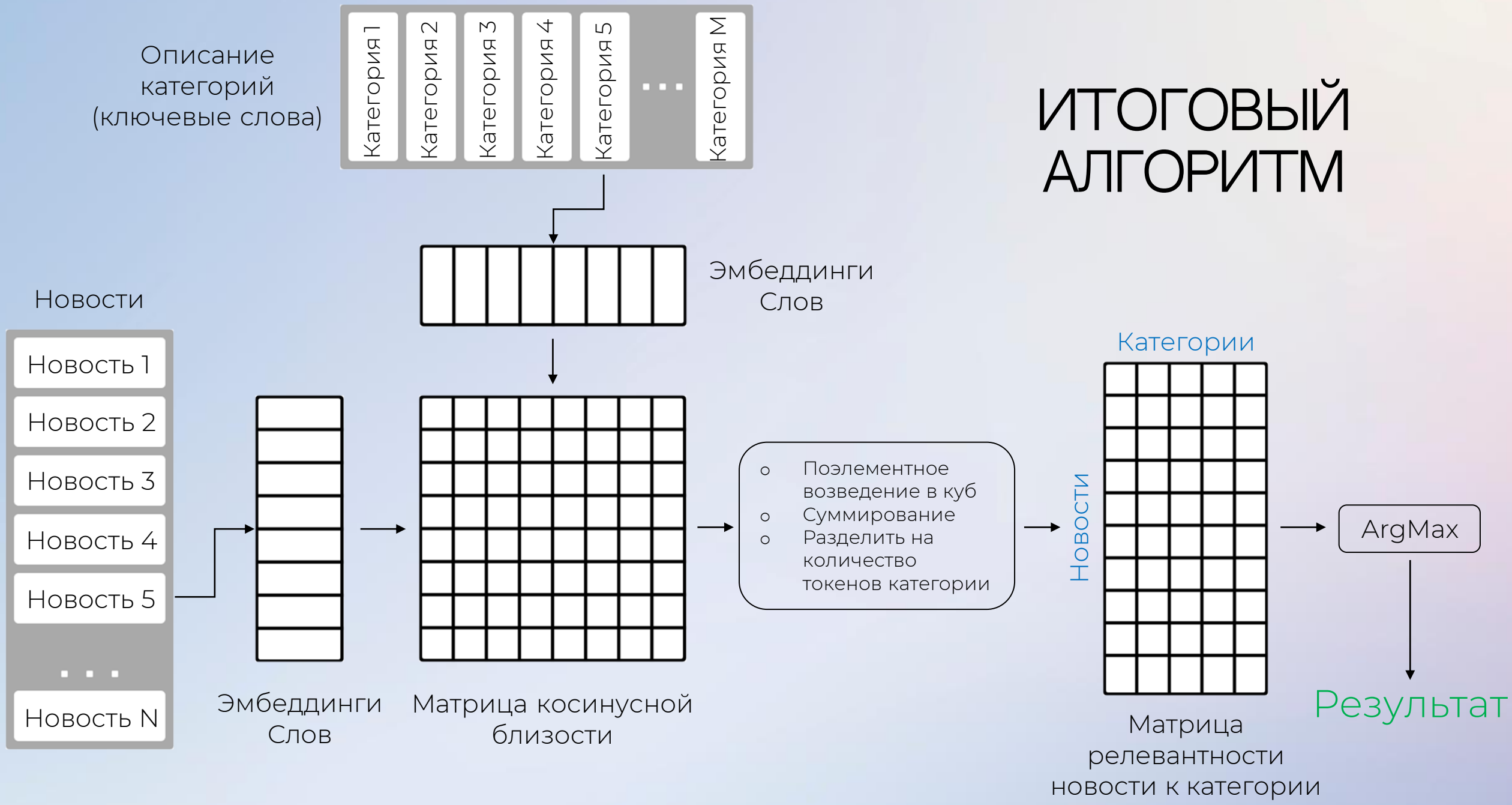
- Разметить все данные. То есть перевести задачу в обучение с учителем. Недостаток: долго или дорого.
- Использовать сторонние датасеты для обогащения данных. Однако обучение на датасете Лента.ру (700к сэмплов) не дало значимых улучшений.

В итоге используем первоначальный датасет и решаем задачу обучения без учителя.

Для процедуры валидации отобрали 1000 случайных новостей из датасета и разметили 3-мя разными людьми. Метрика F1-score измеряется как среднее среди 3-х датасетов.



ИТОГОВЫЙ АЛГОРИТМ



СРАВНЕНИЕ АЛГОРИТМОВ

	model	f1 micro (all)	f1 micro (without rare)	accuracy
0	validation/1. clf_kernel.xlsx	0.295553	0.296112	0.478435
1	validation/2. part_bert_res.xlsx	0.051822	0.052332	0.094283
2	validation/3. closest_word_bert.xlsx	0.032765	0.028534	0.053159
3	validation/4. cats_from_gpt_news.xlsx	0.106653	0.106357	0.199599
4	validation/5. closest_gpt_news.xlsx	0.091608	0.090401	0.178536
5	validation/6. optimazed_clf_kernel.xlsx	0.295553	0.296112	0.478435

ПОИСК ДУБЛИКАТОВ

Недостатки ANN-алгоритмов FAISS, NeoFuzz:

- Пропускает даже явные дубликаты (до 5%)
- Для частотных векторов не хватает памяти
- Эмбеддинги теряют данные и требуют времени на расчет

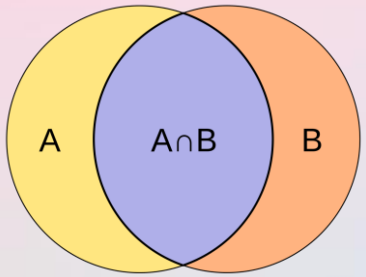
Оптимальное решение – сравнение коэффициента Жаккара по «Bag-of-Words».

Алгоритм показывает хороший результат нахождения дубликатов вкупе с быстрой скоростью (~2 минуты для 50к сэмплов с мультпроцессингом).

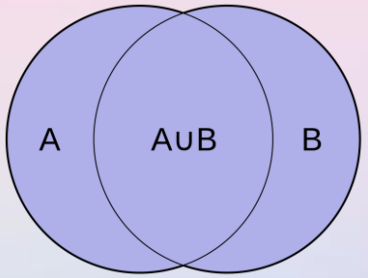
КОЭФФИЦИЕНТ ЖАККАРА

$$K_J = \frac{n(A \cap B)}{n(A \cup B)}$$

Пересечение множеств



Объединение множеств



Пример работы алгоритма:

I. «В лесу упало дерево»
II. «Дерево упало в лесу»
III. «Кошка забежала на дерево»

Лес – 1
Упасть – 2
Дерево – 3
Кошка – 4
Забежать – 5

I. {1, 2, 3}
II. {3, 2, 1}
III. {4, 5, 1}

	I	II	III
I		3/3	1/5
II			1/5
III			

Вывод:
I и II дубликаты

ИНСТРУМЕНТЫ И ТЕХНОЛОГИИ

- Python
- Pandas
- Numpy
- BERT
- Universal Sentence Transformer
- Torch
- Navec
- Faiss
- NeoFuzz
- FastAPI
- Unicorn
- Docker

СПАСИБО
за внимание

GITHUB:

Контакты:

Науменко Дмитрий – https://t.me/naumenko_ds

Кутькина Татьяна – https://t.me/Tatyanna_Kutkina

Дамдинов Зорикто – https://t.me/suzuyajxiii