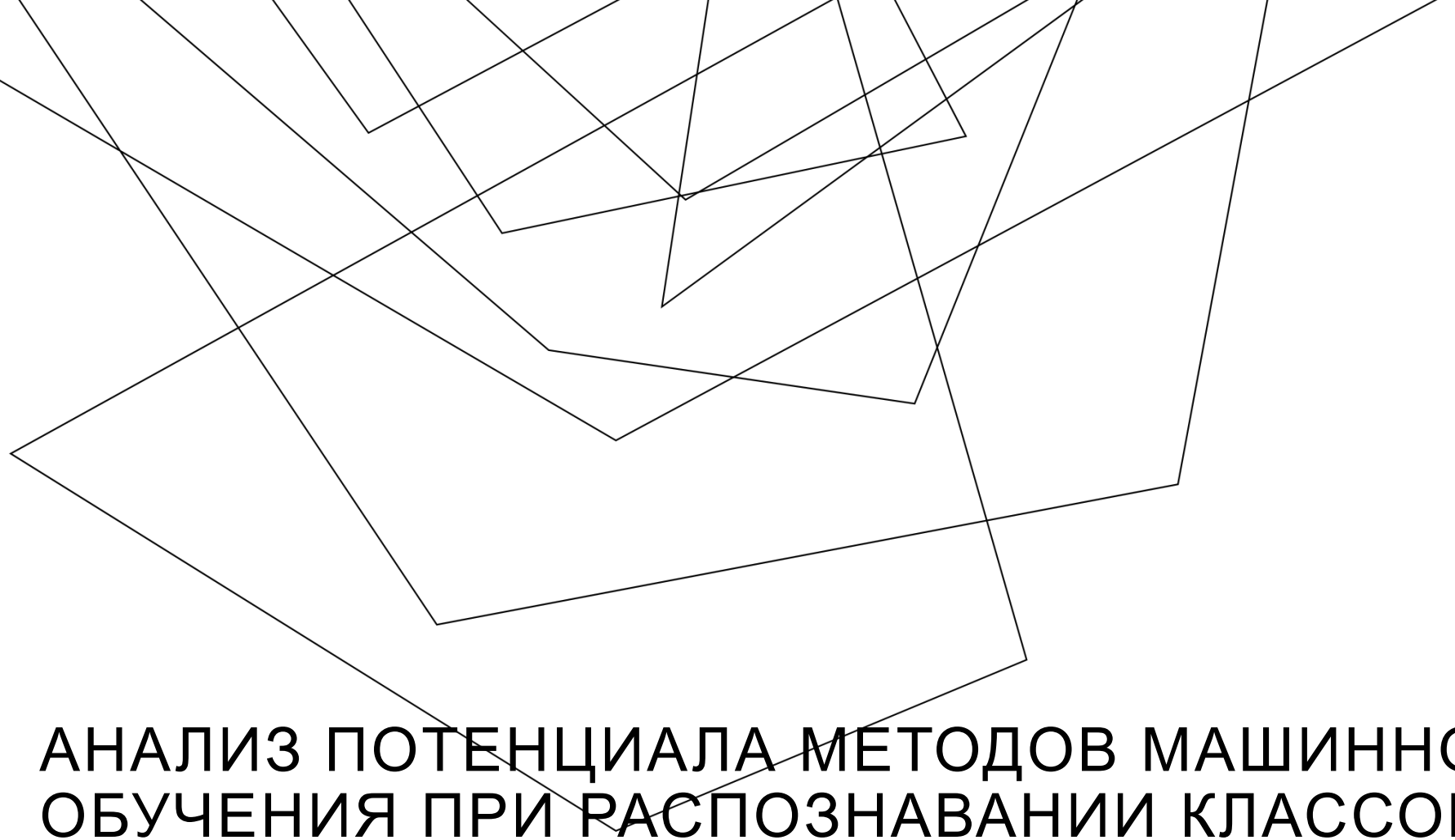


АНАЛИЗ ПОТЕНЦИАЛА МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ПРИ РАСПОЗНАВАНИИ КЛАССОВ В ТРОЙКЕ "ШИЗОФРЕНИЯ-НОРМА-ДЕТСКАЯ РЕЧЬ"

КАРАВАЕВА ОЛЬГА
ДПО КОМП.ЛИНГВИСТИКА НИУ ВШЭ 2023-24

ЗАДАЧА РАБОТЫ

ПРОВЕРИТЬ ВОЗМОЖНОСТЬ ИДЕНТИФИКАЦИИ МЕНТАЛЬНОГО СОСТОЯНИЯ И ВОЗРАСТА ГОВОРЯЩЕГО ПО ТЕКСТОВОЙ ЗАПИСИ РЕЧИ

ПЛАН

1. Бинарная классификация в паре болезнь-норма (tf-idf, LogisticRegression)

2. Мультикласс с включением детской речи (tf-idf, MultinomialNB)

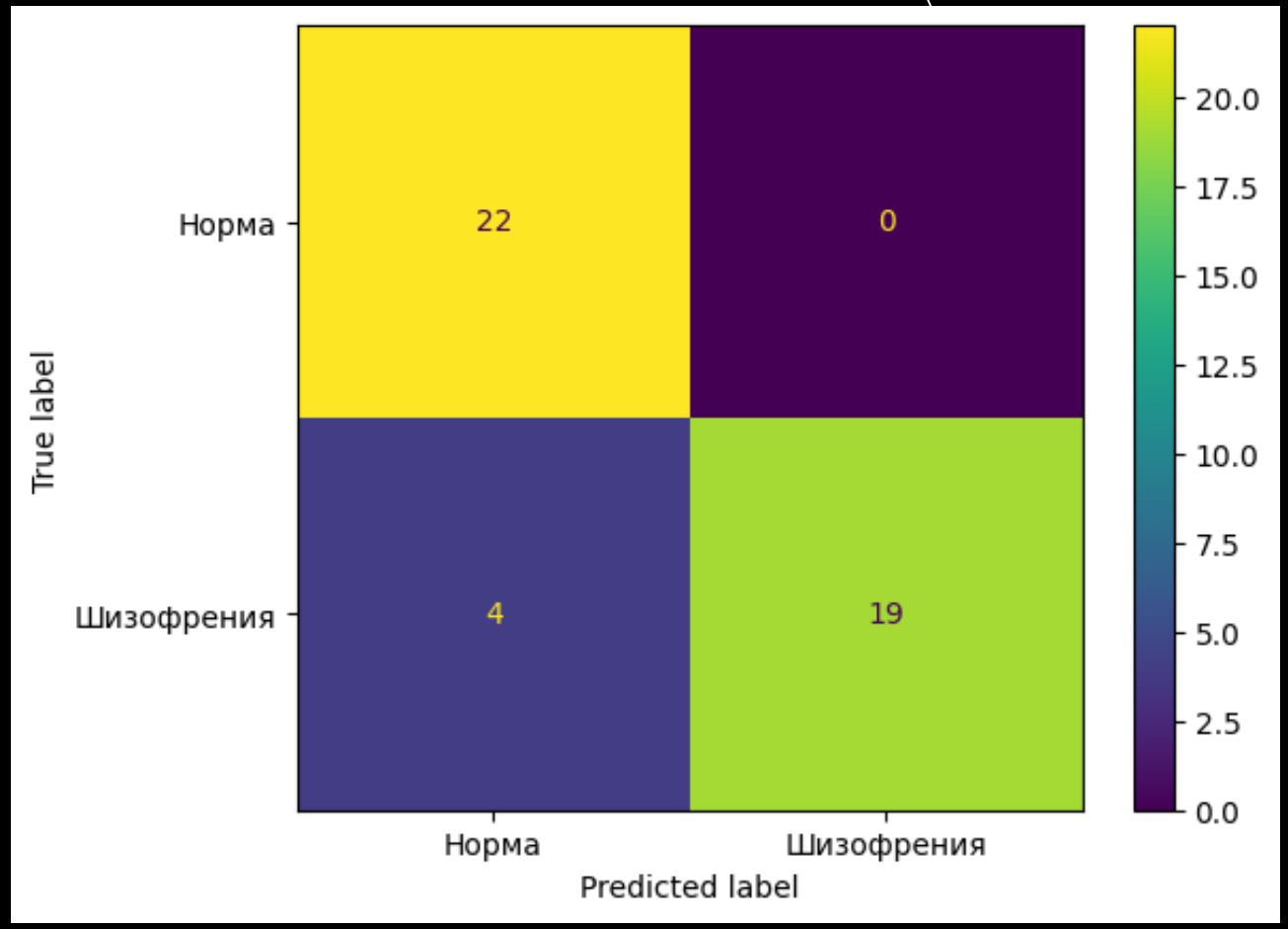
3. Улучшение результатов с помощью

кросс-валидации и ансамблей

4. Влияние объема текста на распознавание классов

5. Визуализация эмбеддингов слов для каждого класса с помощью Word2Vec

РЕЗУЛЬТАТЫ БИНАРНОЙ КЛАССИФИКАЦИИ



	precision	recall	f1-score	support
Норма	0.85	1.00	0.92	22
Шизофрения	1.00	0.83	0.90	23
accuracy			0.91	45
macro avg	0.92	0.91	0.91	45
weighted avg	0.92	0.91	0.91	45

КРИТЕРИИ ВЛИЯНИЯ

Для шизофрении:

- Явления атаксии речи и шизофазии
- Алогизмы или паралогизмы
- Нарушение семантической структуры ассоциаций
- Неологизмы и парафазии
- Нарочито сложный синтаксис, нанизывание грамматических структур
- Многословность

«Ну это не злоба, это остервенение с досадой, которые слегка калькулируют от кофе, от…к воде, которая, знаешь, не настоенная, а отстойная, потому что церемониал земной, подземный и поверхностный, и шахтерный, штольный еще есть. Не кобальды, не землекопы, а улей пчелиный есть, осиный, а есть термитный, а есть муравьиный».

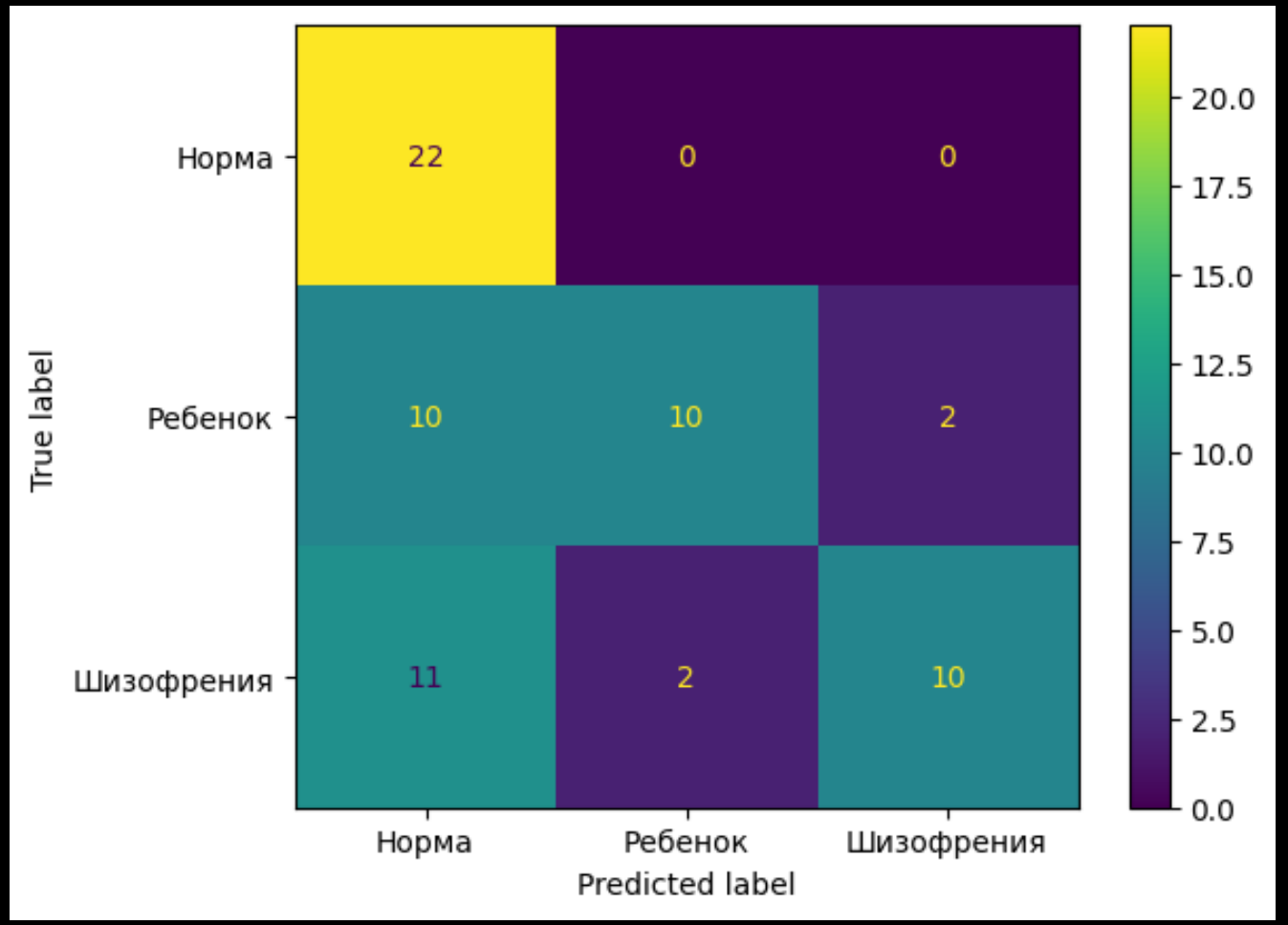
Для детской речи:

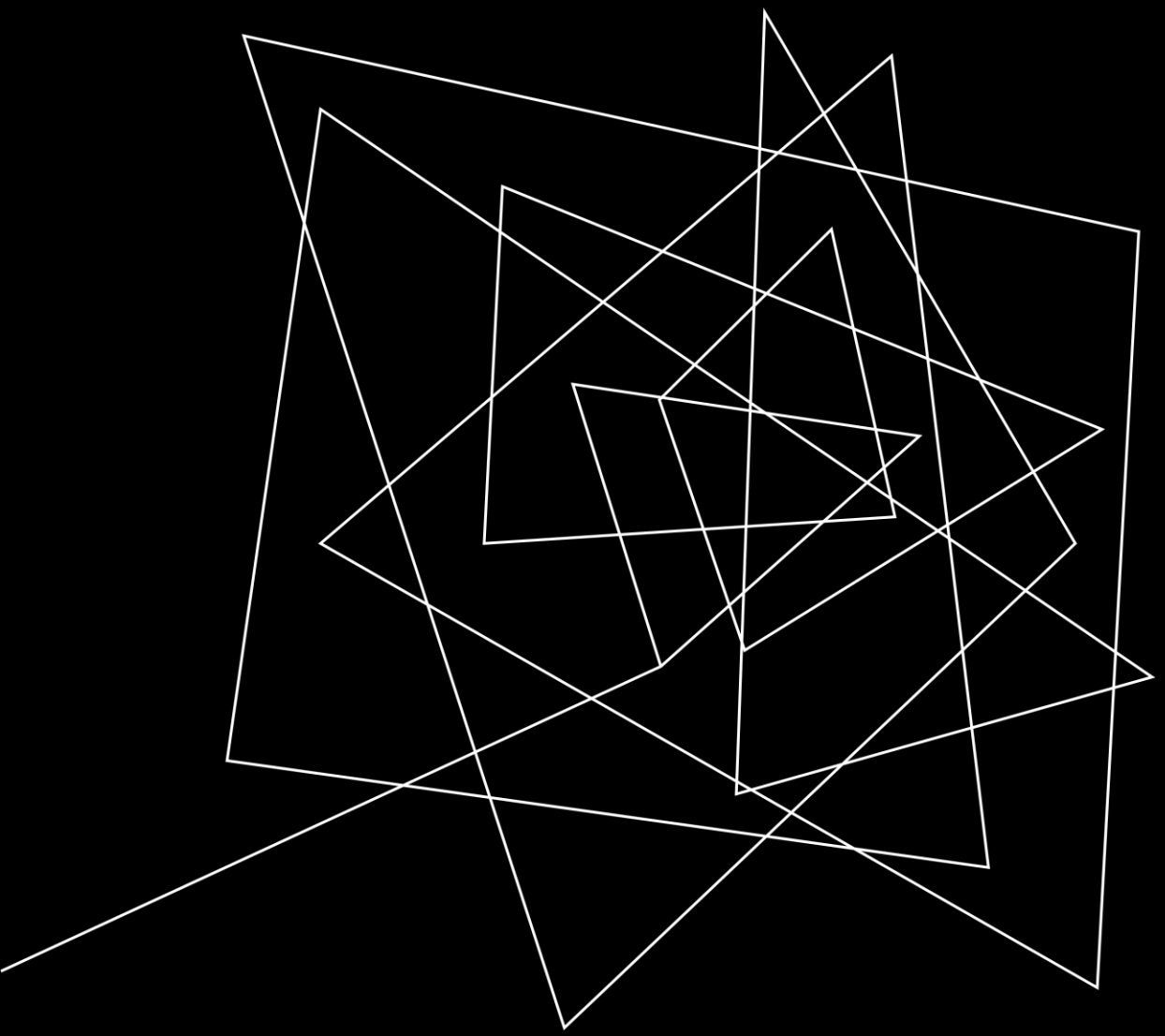
- Скудность словарного запаса в соответствии с возрастом
- Эгоцентрическая речь
- Детское словотворчество
- Примитивность синтаксиса, ошибки
- Короткие фразы в диалоге

«Да, очень хочу. Потому что уже можно делать то, что ты сама захочешь. Еще одна причина, почему я хочу в школу: там я буду без присмотра. Конечно, за мной будет следить учительница, но я смогу сама решать многие вещи. Когда чего-то очень хочется делать, а не разрешают, очень обидно. Например, шоколадку съесть».

РЕЗУЛЬТАТЫ МУЛЬТИКЛАССОВОЙ КЛАССИФИКАЦИИ (БЕЗ K-FOLD)

	precision	recall	f1-score	support
Норма	0.51	1.00	0.68	22
Ребенок	0.83	0.45	0.59	22
Шизофрения	0.83	0.43	0.57	23
accuracy			0.63	67
macro avg	0.73	0.63	0.61	67
weighted avg	0.73	0.63	0.61	67





АНАЛИЗ ОШИБОК

В КЛАСС «НОРМА» ПОДМЕШАНЫ ЛИШНИЕ ДАННЫЕ ОТ ОСТАЛЬНЫХ КЛАССОВ.

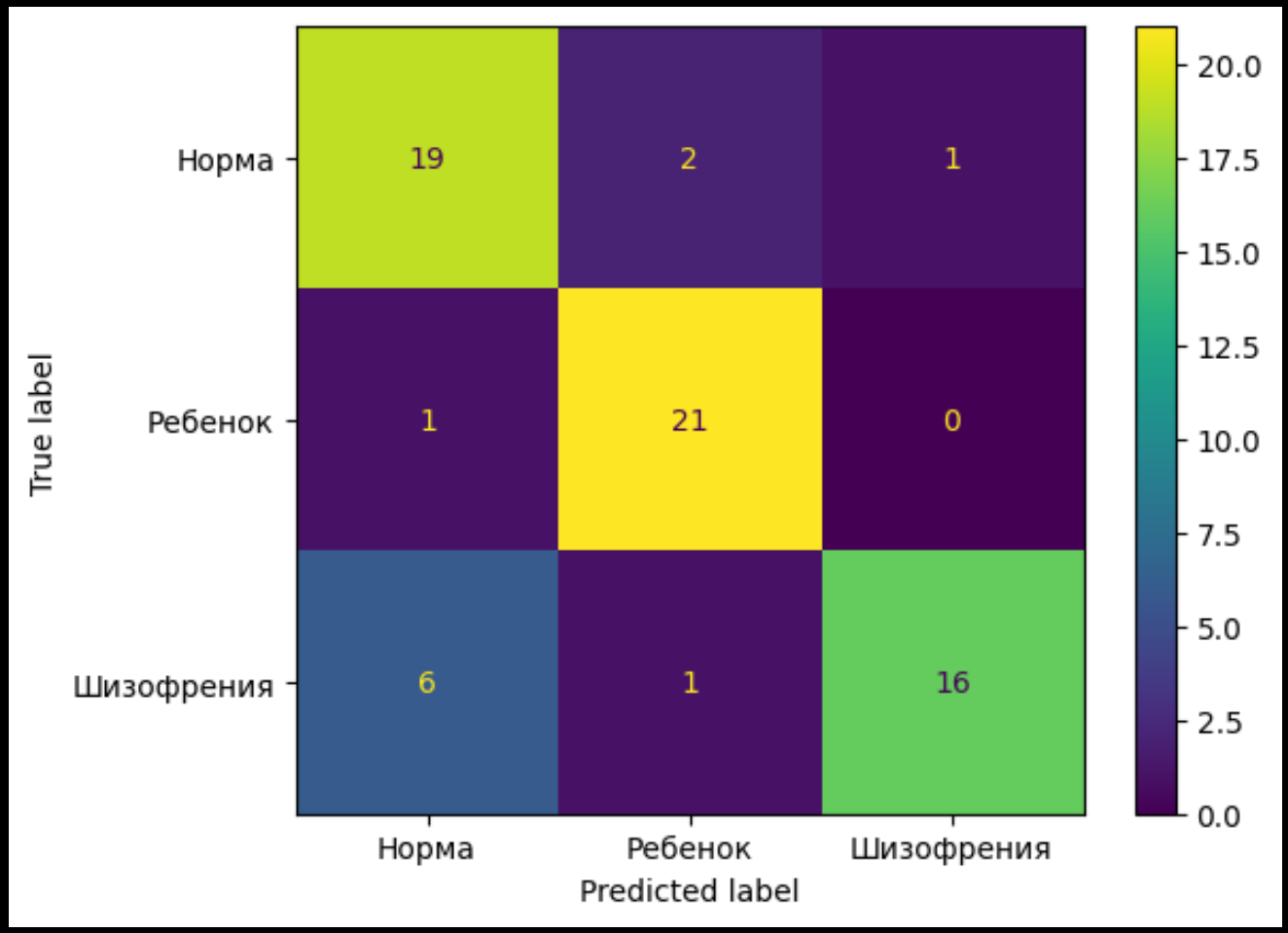
ВОЗМОЖНЫЕ ПРИЧИНЫ:

1. ДАННЫЕ ПО ШИЗОФРЕНИИ И ДЕТСКОЙ РЕЧИ НЕОДНОРОДНЫ (ВОЗРАСТ И СТАДИЯ БОЛЕЗНИ)

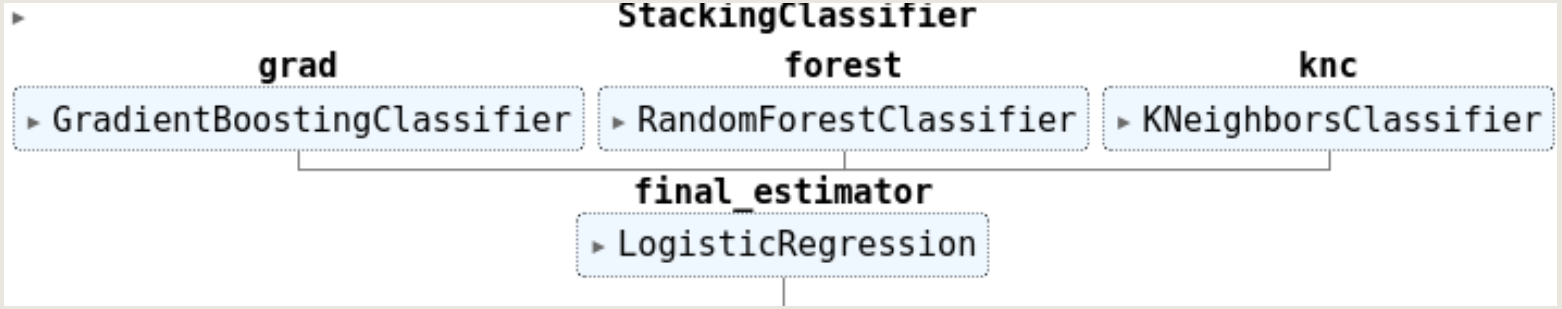
2. НЕ УЧТЕН ОБЪЕМ ТЕКСТОВ

УЛУЧШЕНИЕ РЕЗУЛЬТАТА С ИСПОЛЬЗОВАНИЕМ МЕТОДА K-FOLD

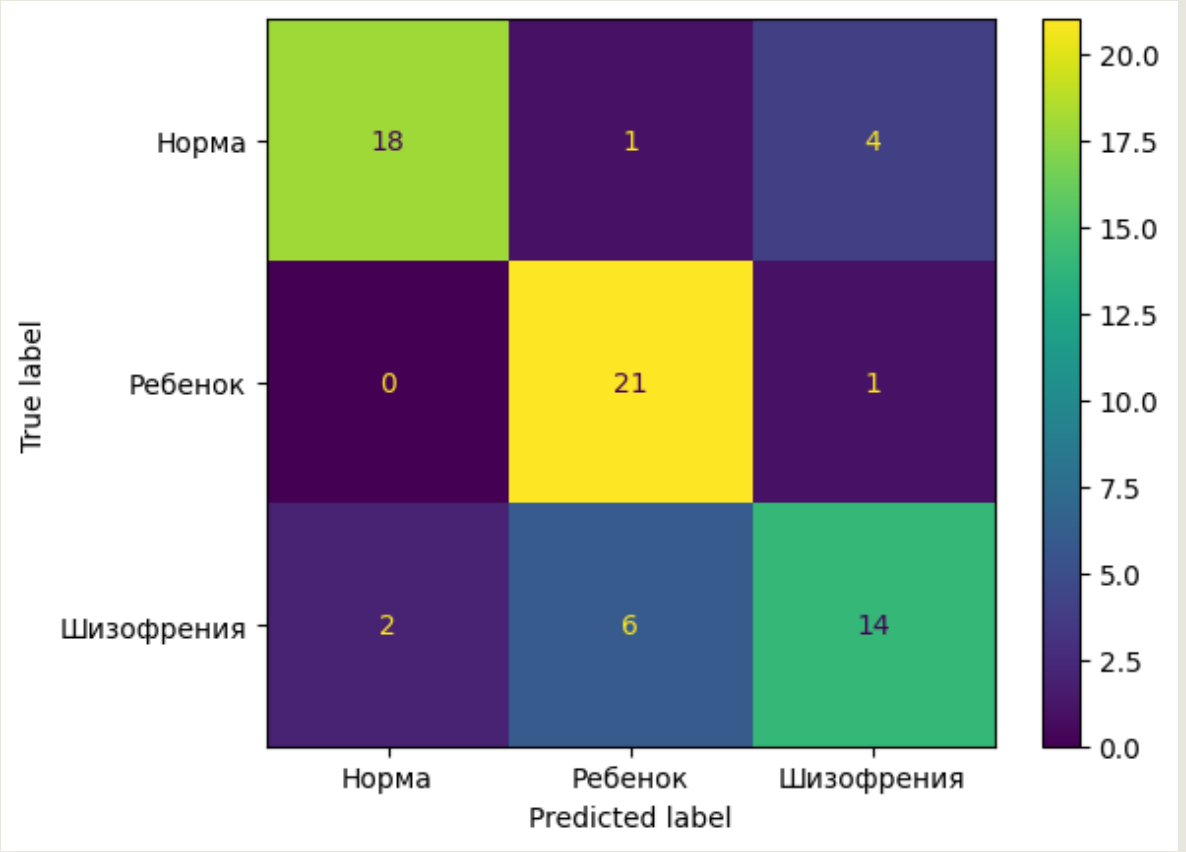
	precision	recall	f1-score	support
Норма	0.73	0.86	0.79	22
Ребенок	0.88	0.95	0.91	22
Шизофрения	0.94	0.70	0.80	23
accuracy			0.84	67
macro avg	0.85	0.84	0.83	67
weighted avg	0.85	0.84	0.83	67



УЛУЧШЕНИЕ РЕЗУЛЬТАТОВ МЕТОДОМ АНСАМБЛЕЙ



	precision	recall	f1-score	support
Норма	0.90	0.78	0.84	23
Ребенок	0.75	0.95	0.84	22
Шизофрения	0.74	0.64	0.68	22
accuracy			0.79	67
macro avg	0.80	0.79	0.79	67
weighted avg	0.80	0.79	0.79	67

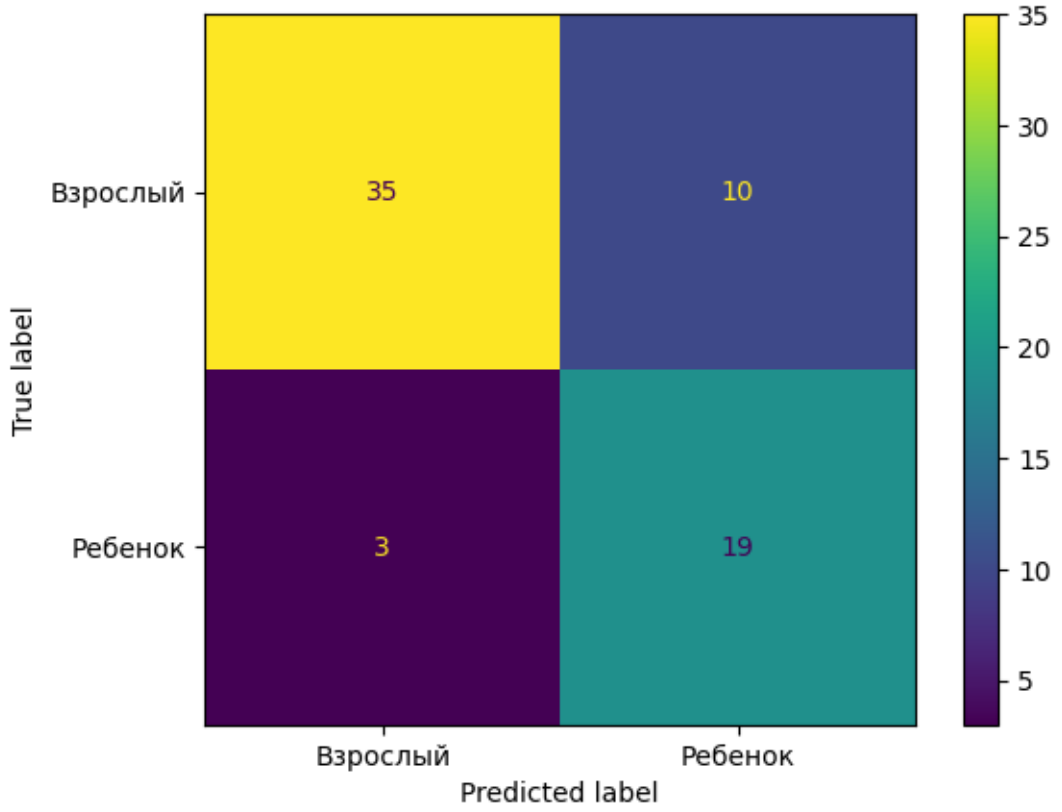


ВЛИЯНИЕ ОБЪЕМА ТЕКСТА НА ВЫБОР ГРУППЫ «ВЗРОСЛЫЙ» (НОРМА + НЕ НОРМА) И «РЕБЕНОК»

1. ТРЕНИРОВОЧНЫЕ ДАННЫЕ БЫЛИ СБАЛАНСИРОВАНЫ RANDOMOVERSAMPLER

2. МОДЕЛЬ ВСЕ РАВНО МНОГО РЕПЛИК РЕБЕНКА ПРИНЯЛА ЗА РЕПЛИКИ ВЗРОСЛОГО

	precision	recall	f1-score	support
Взрослый	0.92	0.78	0.84	45
Ребенок	0.66	0.86	0.75	22
accuracy			0.81	67
macro avg	0.79	0.82	0.79	67
weighted avg	0.83	0.81	0.81	67



ВЫВОДЫ ПО ПРЕОБРАЗОВАНИЯМ

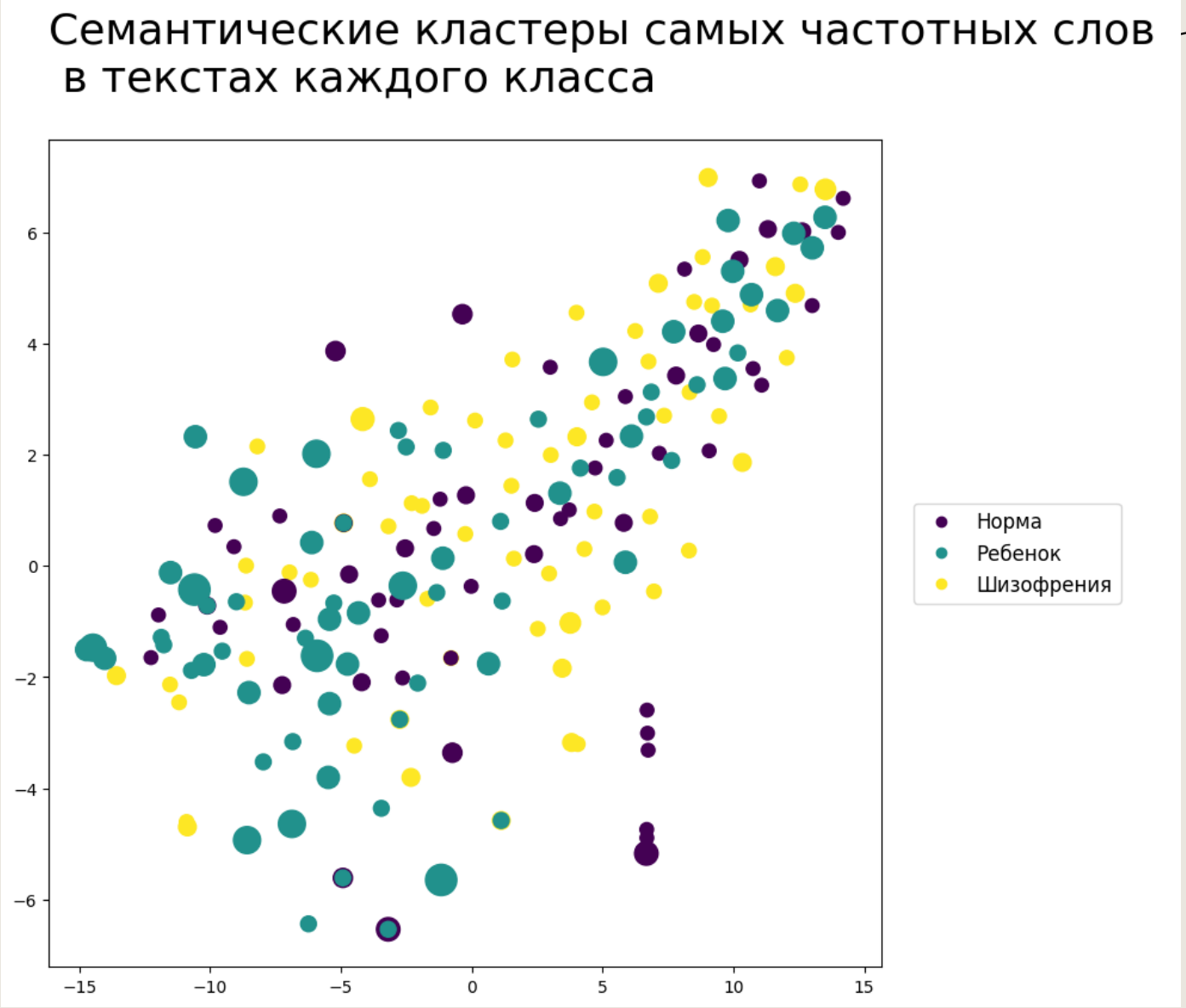
1. Кросс-валидация K-Fold заметно улучшила результаты и «выровняла» их.

2. Ансамбль также поправил результат, но его работа в несколько раз дольше при несколько худшем качестве.

3. Корреляция содержания текста с объемом могла бы помочь улучшить общий результат, но не для любых датасетов.

5. ВИЗУАЛИЗАЦИЯ ЭМБЕДДИНГОВ WORD2VEC

Хорошо визуализировался тезис о лексической бедности детской речи по сравнению со взрослой



СПАСИБО ЗА ВНИМАНИЕ!