УДК 004.934.2
doi:10.31799/1684-8853-2024-4-2-11
EDN: RBUXLJ

Научные статьи
Articles

Подход к распознаванию депрессии по речи человека с использованием полуавтоматической разметки данных

А. Н. Величко[a], *канд. техн. наук, старший научный сотрудник, orcid.org/0000-0002-8503-8512*
А. А. Карпов[a], *доктор техн. наук, профессор, orcid.org/0000-0003-3424-652X, karpov@iias.spb.su*
[a]*Санкт-Петербургский Федеральный исследовательский центр РАН, 14-я линия В. О., 39, Санкт-Петербург, 199178, РФ*

Введение: *с момента выпуска в общий доступ одного из самых больших корпусов, содержащих речь людей с депрессией, Extended Distress Analysis Interview Corpus, в сфере автоматического распознавания речи разработаны новые технологии, применение которых дает возможность повысить качество разметки, а вместе с тем и качество распознавания депрессии.* ***Цель:*** *повысить качество автоматического распознавания депрессии по речи людей с использованием корпуса Extended Distress Analysis Interview Corpus за счет объединения автоматического транскрибирования аудиозаписей с получением временных меток для каждого высказывания, а также экспертной проверки полученных данных для исправления ошибок разметки.* ***Результаты:*** *представлен полуавтоматический подход для разметки аудиоданных с использованием модели Faster-Whisper для текстового транскрибирования речевых записей, набора скриптов для предобработки данных и программного инструментария Praat для ручной проверки полученных транскрипций. В ходе экспериментальных исследований использовано несколько различных методов для решения задач классификации и регрессии. Попытка нормализации данных позволила улучшить значения показателей для метода k-ближайших соседей на предобработанных данных, однако не дала никаких изменений и даже немного ухудшила значения показателей на оригинальных данных. Анализ результатов, полученных в ходе экспериментальных исследований, выявил, что в целом, несмотря на понижение средних значений показателей точности, был сокращен разрыв значений показателей для каждого класса за счет повышения качества распознавания депрессии, что свидетельствует о том, что цель работы достигнута.* ***Практическая значимость:*** *использование представленного подхода позволило улучшить как качество разметки, так и качество автоматического распознавания депрессии.* ***Обсуждение:*** *в дальнейшем планируется использовать полученную разметку для проведения экспериментальных исследований при создании метода многомодального распознавания депрессии человека по аудио, видео и текстовым данным.*

Ключевые слова *— анализ речи, речевые технологии, компьютерная паралингвистика, деструктивные явления, предобработка данных, автоматическое распознавание депрессии по речи.*

Для цитирования: Величко А. Н., Карпов А. А. Подход к распознаванию депрессии по речи человека с использованием полуавтоматической разметки данных. *Информационно-управляющие системы*, 2024, № 4, с. 2–11. doi:10.31799/1684-8853-2024-4-2-11, EDN: RBUXLJ

For citation: Velichko A. N., Karpov A. A. An approach to depression detection in speech using a semi-automatic data annotation. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2024, no. 4, pp. 2–11 (In Russian). doi:10.31799/1684-8853-2024-4-2-11, EDN: RBUXLJ

Введение

Большое депрессивное расстройство (БДР) является распространенным аффективным расстройством, которое характеризуется подавленным настроением, потерей интереса или утратой способности получать удовольствие от приятной ранее деятельности, расстройством сна и (или) аппетита, психомоторным возбуждением или заторможенностью, повышенной утомляемостью и снижением энергичности, снижением самооценки или неадекватным чувством вины, снижением способности к концентрации внимания или заторможенным мышлением, суицидальными тенденциями. Данное состояние приводит к ухудшению жизнедеятельности, может стать причиной инвалидности [1].

Специалист при беседе с пациентом учитывает особенности речевой продукции при затрагивании значимых для пациента тем. В ряде межкультурных исследований [2–4] показано, что проявления депрессии схожи во многих культурах, что в перспективе позволит использовать автоматические системы распознавания депрессии по речи как минимум для родственных языковых групп вне зависимости от языка данных для обучения. Современные методы машинного и глубокого обучения могут обрабатывать многомодальную информацию (видео, аудио, текст и иные модальности), что дает возможность эффективно отличать человека с БДР от здорового человека. Ряд работ [5–8] посвящен определению психического состояния пользователей социальных сетей на основе контента, который они публикуют, что является одним из возможных практических применений систем распознавания депрессии. Многие аналитические работы [9–11] рассматривают теоретические и практические исследования, в которых представлены существующие на момент написания многомо-

дальные и одномодальные корпуса, содержащие речь людей с депрессией, а также систематизированные признаки депрессии по модальностям. Несмотря на актуальность тематики, в настоящее время количество открытых речевых и многомодальных корпусов, содержащих речь людей с установленной депрессией, ограничено. При этом существующие корпуса имеют такие недостатки, как малое количество данных, дисбаланс классов в данных, неточность разметки и пр. Одним из самых больших корпусов на сегодня является корпус Extended Distress Analysis Interview Corpus (E-DAIC) [12]. Различные подходы к предобработке данных на корпусе E-DAIC сравниваются в работе [12], а сравнительный анализ автоматической и полуавтоматической, т. е. с участием человека, предобработки текста проведен в исследовании [13]. Можно заметить, что, несмотря на технологический прогресс, во многих случаях качество распознавания наличия депрессии оказывается выше, если была проведена полуавтоматическая предобработка.

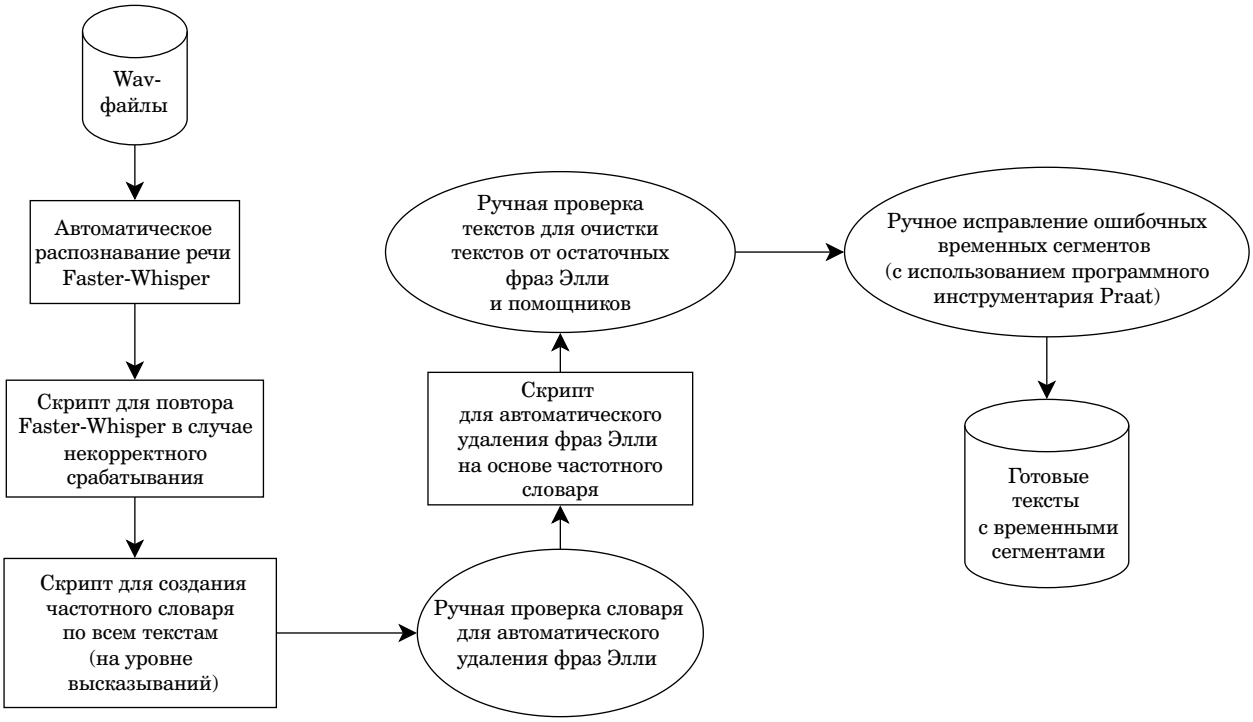
На данный момент большинство работ на корпусе E-DAIC посвящено задаче регрессии, т. е. определению степени тяжести депрессии на основе шкалы PHQ-8. Задача регрессии с использованием этого корпуса представлена на соревнованиях по аудиовизуальному распознаванию эмоций AVEC-2019 [12]. В нашей работе проведены экспериментальные исследования как для задачи регрессии, так и для задачи бинарной классификации (определения наличия или отсутствия депрессии).

Подход к предобработке речевых данных

В качестве исследовательских данных для обучения и тестирования предложенного подхода выбран многомодальный корпус клинических интервью E-DAIC. «Корпус разработан для симуляции стандартных процессов определения того, имеется ли у человека риск ПТСР (посттравматического стрессового расстройства) и БДР» [14]. Он содержит 275 интервью в двух режимах: «автоматическое интервью — интервью проводилось в автоматическом режиме с Элли; «Волшебник Оз» (Wizard of Oz, WoZ) или «Гудвин» — интервью проводилось анимированным виртуальным интервьюером по имени Элли, которую контролировал интервьюер в другой комнате» [14]. Подробное описание корпуса, процесса сбора данных и обнаруженных авторами закономерностей проявлений ПТСР и БДР в речи, помимо оригинальных работ, представлено в работах [14, 15]. Данный корпус является частью корпуса DAIC и расширенной версией корпуса DAIC-WoZ [16], а разметка данных сформирована автоматически с использованием методов автоматического распознавания речи.

В ходе анализа корпуса E-DAIC был выявлен ряд недостатков, потенциально снижающих показатели качества распознавания моделей. Среди них: случаи, когда один участник перебивает другого; некорректная разметка временных сегментов в транскрипции и ее несовпадение с аудио; фразы интервьюера Элли и помощников в транскрипциях; несовпадения в целевых значениях классов, которые также были замечены другими исследователями в корпусе DAIC-WoZ [17, 18]. Ввиду того, что версия корпуса E-DAIC была размечена с использованием автоматических инструментов, существовавших на момент публикации корпуса в 2019 г., качество транскрипций и точность временных сегментов ниже, чем это возможно сделать с использованием современных методов. Поэтому целью данной работы является улучшение качества распознавания депрессии на корпусе E-DAIC за счет объединения автоматического транскрибирования (распознавания) аудиозаписей с получением временных сегментов для каждого высказывания, а также ручной проверки полученных данных и, при необходимости, исправления ошибок. Последовательность действий при предобработке данных представлена на рис. 1.

Для транскрибирования речи были опробованы различные модели Whisper [19] и Faster-Whisper (https://github.com/SYSTRAN/faster-whisper). Наилучшей моделью оказалась самая большая модель Faster-Whisper — large-v3 — благодаря высоким показателям скорости, затрат по вычислительным ресурсам и качества транскрипций. В отличие от оригинальных моделей Whisper, Faster-Whisper обладает преимуществом ввиду меньшего потребления вычислительных ресурсов и более высокой скорости обработки данных. Фактически она является надстройкой над оригинальным решением Whisper, а потому качество транскрипций соответствует результатам оригинальной модели. Необходимость повторения скрипта для запуска Faster-Whisper обусловлена тем, что в некоторых случаях модель срабатывает некорректно и создает повторы фразы, если не уверена в том, что верно распознала фразу. Решением этого феномена на данный момент является настройка температуры модели (гиперпараметр, значения которого находятся между 0 и 1; чем выше его значение, тем более случайным будет вывод, и наоборот, чем ниже значение, тем вывод более детерминированный; если значение равно 0, то модель будет использовать логарифмическую вероятность для автоматического повышения температуры до тех пор, пока не будут достиг-



■ ***Рис. 1.*** Этапы полуавтоматической предобработки данных
■ ***Fig. 1.*** Semi-automatic data preprocessing steps

нуты определенные пороговые значения), а также ее повторный запуск, так как некорректные срабатывания происходят не при каждом запуске (ввиду специфики интервью три повторения подряд одной фразы является максимально возможным количеством повторений).

Произведена попытка автоматической диаризации речи дикторов с использованием Pyannote (версий v2 и v3) [20] как наиболее качественного подхода на сегодня. К сожалению, при использовании данного инструмента получено неудовлетворительное качество диаризации в случаях, когда на записи присутствуют три различных женских голоса (участника, помощника и интервьюера Элли).

Было решено произвести полуавтоматическую разметку текстов речевых транскрипций с использованием скриптов и ручной разметки. Первым делом необходимо было использовать скрипт для создания частотного словаря на уровне высказываний по всем текстам. Данный словарь был проверен вручную, отобраны фразы, которые точно принадлежат Элли. На следующем шаге эти фразы были автоматически вырезаны из текстов транскрипций. Наиболее время- и трудозатратным шагом явилось ручное удаление остаточных фраз Элли и помощников. Данный шаг был необходим с тем, чтобы в транскрипциях остались только фразы участников. Ввиду того, что в транскрипциях встречались ошибочные временные сегменты, необходимо было их исправить, что сделано с использованием программного инструментария Praat [21]. Эти сегменты были представлены отрезками диалогов, когда в одном временном сегменте в транскрипции показана транскрипция речи и участников, и Элли. Они являлись результатом автоматического распознавания речи в тех частях диалогов, где либо происходило наложение речи участников и Элли, либо ответ следовал моментально за вопросом, без пауз.

Таким образом, среди очевидных ошибок (видимых человеческому глазу при прочтении текстовых транскрипций и одновременном прослушивании аудио) в автоматическом распознавании речи можно выделить:

— ошибки во временных сегментах (отсутствие пауз между фразами участника);

— ошибки в автоматическом распознавании (отсутствие некоторых фраз участника; наличие фраз интервьюера Элли или помощника во фразах участника ввиду некорректного распознавания временных сегментов; ошибки распознавания фраз участника ввиду либо наложения речи интервьюера Элли или помощника и участника, либо слишком тихо произнесенной фразы).

Поскольку в некоторых работах найдены несоответствия в целевых значениях классов (присвоение класса 0 — «отсутствие депрессии» объектам класса 1 — «наличие депрессии»), про-

ведена также проверка соответствия бинарных значений классов шкале PHQ-8. Согласно данной шкале, если участник набирает 10 и более баллов, такой результат расценивается как выраженная депрессия, а при результате 20 и более баллов — тяжелая депрессия. Так, в корпусе E-DAIC обнаружено 20 несоответствий данной шкале (13 в обучающем наборе, 3 — в отладочном, 4 — в тестовом), при том, что в корпусе DAIC-WoZ одно такое несоответствие. Данные несоответствия целевых значений исправлены вручную, так как они могли заметно повлиять на результаты обучения моделей и автоматического распознавания депрессии. Было получено 189 аудиозаписей в корпусе, принадлежащих классу 0, а классу 1 — 86.

Также обнаружены аудиозаписи, в которых количество каналов и частота дискретизации отличались от большинства аудиозаписей. В корпусе значительная часть аудиозаписей имеет один канал записи (моно), частота дискретизации — 16 кГц, однако было обнаружено 33 аудиозаписи, имеющие два канала (стерео), частоту дискретизации 48 кГц. Путем конвертации эти аудиозаписи приведены к общим параметрам, поскольку такое различие могло повлиять на процесс и качество вычисления акустических признаков.

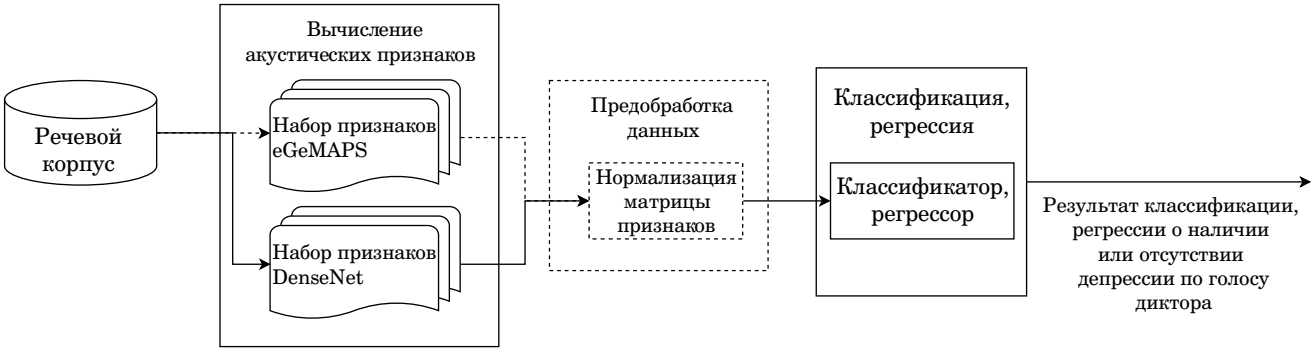
Экспериментальные исследования и полученные результаты

Экспериментальные исследования проведены с использованием предложенного метода распознавания депрессии по речи (рис. 2). На основе работы [14] были вычислены два набора акустических признаков с помощью программного инструментария openSMILE (http://audeering.com/technology/opensmile/): 88 признаков eGeMAPS [22] и 1920 признаков DenseNet [23]. Процесс вычисления наборов акустических признаков, сами признаки и используемая для экспериментальных исследований нейросетевая архитектура TabNet [24] подробно рассмотрены в работе [14].

В нашей работе для бинарной задачи определения депрессии (присутствует заболевание у говорящего или нет) были применены и исследованы несколько различных машинных классификаторов: глубокая нейронная сеть TabNet, градиентный бустинг CatboostClassifier, метод линейного дискриминантного анализа (Linear Discriminant Analysis, LDA), метод k-ближайших соседей (k-Nearest Neighbours, k-NN). Данные классификаторы выбраны с учетом того, что их обучение занимает малое время относительно многих других подходов и не требует большого количества вычислительных ресурсов.

Наборы данных для проведения экспериментальных исследований состоят из 163 экземпляров записей для обучения (train), других 56 — для валидации (val) и еще 56 — для тестирования (test). Набор для валидации (val) применялся в случаях, когда классификатор имел встроенную возможность его использовать (TabNet, Catboost). У классических методов машинного обучения (LDA, k-NN) такой возможности нет, поэтому для работы с ними набор для валидации был объединен с набором для обучения (train+val). Объективное тестирование проводилось, соответственно, на тестовом наборе данных (test).

Для сравнения результатов предобработки данных экспериментальные исследования выполнялись с использованием двух наборов данных: оригинального набора (описанные выше изменения не были внесены, он использовался в том виде, в котором представлен авторами) и предобработанного набора (с внесенными изменениями, описанными выше). Количественные результаты экспериментальных исследований в задаче бинарной (наличие или отсутствие депрессии) классификации депрессии представлены в табл. 1. В качестве количественных пока-



■ ***Рис. 2.*** Обобщенная схема метода распознавания депрессии по речи
■ ***Fig. 2.*** A unified scheme of the method for depression detection by speech

зателей для оценивания методов выбраны: взвешенная средняя F-мера (F1 WA), F-мера для каждого класса, невзвешенная средняя F-мера (F1 UA), взвешенная средняя полнота (WAR), полноты для каждого класса (Recall), а также невзвешенная средняя полнота (Unweighted Average Recall, UAR). Классы Depressed (1 — «наличие депрессии») и Non-Depressed (0 — «отсутствие депрессии») обозначены D и ND соответственно.

Гиперпараметры машинных классификаторов использованы преимущественно базовые, но были изменены параметры конфигурации запуска TabNet batch_size = 64 (по умолчанию 1024) и virtual_batch_size = 32 (по умолчанию 128). В метод Catboost были внесены изменения в целях построения метода классического градиентного бустинга и использования его особенностей: параметр весов классов (1-му классу присвоен вес в два раза больше, чем 0-му, так как в данных присутствует дисбаланс), установлен детектор переобучения, выставлена схема классического градиентного бустинга boosting_type='Plain'. Количество соседей для метода k-NN равно трем.

Несмотря на то, что некоторые значения показателей на оригинальных данных выше (см. табл. 1), на предобработанных данных разброс между значениями показателей для каждого класса гораздо меньше. Также можно заметить тенденцию, что при добавлении набора признаков eGeMAPS существенно снижаются значения показателей для класса D, но при этом повышаются значения для класса ND. Попытка нормализации данных при использовании метода k-NN позволила улучшить значения показателей на предобработанных данных, однако не дала никаких изменений и даже немного ухудшила значения показателей на оригинальных данных.

Результаты экспериментальных исследований в регрессионной задаче распознавания депрессии представлены в табл. 2. Применены следующие методы регрессии для набора признаков DenseNet: CatboostRegressor, логистическая и линейная регрессии (Logistic and Linear Regression). В качестве показателей для оценивания методов использованы коэффициент корреляции согласования (Concordance Correlation

■ ***Таблица 1.*** Результаты экспериментальных исследований методов для бинарной классификации депрессии на предобработанных и оригинальных данных

■ ***Table 1.*** Depression classification results on both preprocessed and original data

Метод	Вариант разбиения данных	F1 WA	F1 UA	F1 (ND)	F1 (D)	WAR	UAR	Recall (ND)	Recall (D)
Предобработанные данные									
TabNetClassifier (DenseNet)	train/val/test	0,55	0,53	0,65	0,41	0,56	0,53	0,67	0,39
TabNetClassifier (DenseNet + eGeMAPS)	train/val/test	0,54	0,49	**0,73**	0,25	0,60	0,53	**0,89**	0,17
CatboostClassifier (DenseNet)	train/val/test	**0,61**	**0,61**	0,69	**0,48**	**0,61**	**0,59**	0,72	**0,45**
CatboostClassifier (DenseNet + eGeMAPS)	train/val/test	0,51	0,45	0,72	0,19	0,58	0,51	**0,89**	0,12
LDA (DenseNet)	train+val/test	0,52	0,48	0,68	0,28	0,56	0,50	0,78	0,22
k-NN (DenseNet)	train+val/test	0,52	0,48	0,66	0,31	0,54	0,50	0,73	0,26
k-NN (DenseNet, L2-norm)	train+val/test	0,53	0,49	0,66	0,32	0,55	0,50	0,74	0,27
Оригинальные данные									
TabNetClassifier (DenseNet)	train/val/test	0,59	0,50	0,71	**0,28**	0,59	0,50	0,70	**0,30**
TabNetClassifier (DenseNet + eGeMAPS)	train/val/test	0,60	0,45	0,78	0,12	0,65	0,47	0,87	0,08
CatboostClassifier (DenseNet)	train/val/test	0,63	0,50	0,79	0,22	0,67	0,51	0,86	0,16
CatboostClassifier (DenseNet + eGeMAPS)	train/val/test	0,61	0,49	0,76	0,22	0,63	0,49	0,80	0,19
LDA (DenseNet)	train+val/test	**0,64**	0,49	**0,82**	0,16	**0,70**	**0,52**	**0,93**	0,10
k-NN (DenseNet)	train+val/test	0,63	**0,51**	0,78	0,23	0,66	**0,52**	0,84	0,19
k-NN (DenseNet, L2-norm)	train+val/test	0,63	0,50	0,78	0,23	0,66	0,51	0,84	0,18

■ ***Таблица 2.*** Результаты экспериментальных исследований в регрессионной задаче распознавания депрессии на предобработанных и оригинальных данных

■ ***Table 2.*** Depression regression results on both preprocessed and original data

Метод (с использованием набора признаков DenseNet)	Вариант разбиения данных	CCC (test) ↑	RMSE (val/test) ↓
Предобработанные данные			
CatboostRegressor	train/val/test	0,002	5,69 / **6,82**
LogisticRegression	train+val/test	0,006	– / 9,14
LinearRegression	train+val/test	**0,047**	– / 6,95
Оригинальные данные			
CatboostRegressor	train/val/test	0,003	5,50 / **6,25**
LogisticRegression	train+val/test	**0,085**	– / 8,53
LinearRegression	train+val/test	0,061	– / 6,64

■ ***Таблица 3.*** Результаты сравнения экспериментальных исследований с аналогами в регрессионной задаче распознавания депрессии

■ ***Table 3.*** Comparison of experimental results in depression regression recognition

Метод	Признак, классификатор	CCC ↑		RMSE ↓	
		val	test	val	test
Ringeval F., et al. [12]	MFCCs, GRU	0,198	–	7,28	–
	eGeMAPS, GRU	0,076	–	7,78	–
	BoAW-M, GRU	0,102	–	6,32	–
	BoAW-e, GRU	0,272	0,045	6,43	8,19
	DenseNet, GRU	0,165	–	8,09	–
	VGG, GRU	0,305	0,108	8,00	9,33
Ray A., et al. [25]	Funct MFCC, BLSTM	–	–	**5,11**	–
	Funct eGeMAPS, BLSTM	–	–	5,52	–
	BOAW-M, BLSTM	–	–	5,66	–
	BoAW-e, BLSTM	–	–	5,50	–
	DenseNet, BLSTM	–	–	5,65	–
Makiuchi M. R., et al. [26]	VGG, CNN	**0,338**	**0,199**	5,97	7,02
	VGG, GCNN-LSTM	0,497	–	5,70	–
Метод с использованием предобработанных данных	DenseNet, LinearRegression	–	0,047	–	**6,95**
Метод с использованием оригинальных данных	DenseNet, LogisticRegression	–	0,085	–	8,53

Coefficient, CCC) и среднеквадратичная ошибка (Root Mean Squared Error, RMSE).

По приведенным значениям видно, что результаты CCC и RMSE на оригинальных данных лучше, чем на предобработанных. Наша гипотеза заключается в том, что здесь работает та же логика, что и при классификации депрессии, где значения средних показателей были выше в результатах для оригинальных данных. При этом значения показателей для каждого класса более сбалансированы в результатах классификации на предобработанных данных. Для проверки этой гипотезы необходимо вычислить дополнительно показатели CCC и RMSE для регрессионной задачи распознавания депрессии, которые будут учитывать значения показателей для каждого класса.

Сравнение с аналогами в задаче регрессии представлено в табл. 3. Полученные результаты RMSE для оригинальных и предобработанных данных оказались на уровне с работой организаторов, представленной на соревнованиях AVEC-2019.

Заключение

В статье представлен полуавтоматический подход к предобработке данных на корпусе E-DAIC с использованием Faster-Whisper для текстового транскрибирования речи и программного инструментария Praat для ручной проверки полученных транскрипций. Проведены экспериментальные исследования с несколькими методами распознавания депрессии для задач бинарной классификации и регрессии. На основе полученных количественных результатов можно заключить, что цель работы, а именно повышение качества автоматического распознавания депрессии по речи человека, достигнута. Несмотря на понижение средних значений показателей, сокращен разрыв значений показателей для каждого класса за счет повышения качества распознавания класса Depressed (1 — «наличие депрессии»).

Среди ограничений, которые были преодолены в ходе работы, можно выделить: небольшое количество данных в корпусе (по сравнению с корпусами, предназначенными для других задач, например, автоматического распознавания речи, распознавания эмоций в речи и пр.); дисбаланс классов в данных более чем в два раза (аудиозаписей в корпусе, принадлежащих классу 0 «отсутствие депрессии», — 189, а классу 1 «наличие депрессии», — 86); ограничения, связанные с возможностями современных методов распознавания речи и диаризации. Ограничение, связанное с дисбалансом классов в данных, было преодолено путем придания объектам минорного класса большего веса в классификаторе CatboostClassifier, а ограничение, связанное с возможностями современных методов распознавания речи и диаризации, — путем полуавтоматической предобработки данных. Проблему небольшого количества данных в корпусе теоретически можно решить, объединив несколько корпусов (что, однако, не исключает зашумления данных ввиду различных условий их записи), либо путем аугментации данных (что также несет в себе риски зашумления данных).

Практическое применение предложенного метода тоже имеет ограничения, среди которых относительно низкое качество распознавания депрессии (повышается путем комплексирования предложенного метода) и необходимость участия специалистов для проведения корректировки на первичных этапах внедрения. Отдельно стоит отметить, что предложенный метод может быть использован в качестве вспомогательного средства, а не самостоятельного, ввиду описанных выше ограничений.

Полученные в ходе предобработки аудиоданных корпуса транскрипции планируется использовать в ходе дальнейших исследований на базе корпуса E-DAIC для многомодального распознавания депрессии, а также для многозадачных систем определения различных паралингвистических явлений и аффективных состояний по разговорной речи людей [15, 27]. Такие системы потенциально могут быть полезны при телеконсультировании в первичном звене здравоохранения для скрининга тревожно-депрессивных расстройств ввиду необходимости обработки многомодальной информации в условиях удаленного консультирования и лечения специалистами [28].

Финансовая поддержка

Работа выполнена при финансовой поддержке Российского научного фонда (проект № 22-11-00321, https://www.rscf.ru/project/22-11-00321/).

Литература

1. *Depressive disorder WHO (depression).* WHO. 2023. https://www.who.int/news-room/factsheets/detail/depression (дата обращения: 12.06.2024).
2. **Singer K.** Depressive disorders from a transcultural perspective. *Social Science & Medicine*, 1975, no. 9, pp. 289–301. doi:10.1016/0037-7856(75)90001-3
3. **Alghowinem Sh., Goecke R., Epps J., Wagner M., Cohn J.** Cross-cultural depression recognition from vocal biomarkers. *Proc. of INTERSPEECH-2016*, San Francisco, USA, 2016, pp. 1943–1947. doi:10.21437/Interspeech.2016-1339
4. **Kaya H., Fedotov D., Dresvyanskiy D., Doyran M., Mamontov D., Markitantov M., Salah A., Kavcar E., Karpov A., Salah A.** Predicting depression and emotions in the crossroads of cultures, paralinguistics, and non-linguistics. *Proc. of the 9th ACM Intern. Workshop on Audio/visual Emotion Challenge (AVEC '19)*, Nice, France, 2019, pp. 27–35. doi:10.1145/3347320.3357691
5. **Браницкий А. А., Шарма Я. Д., Котенко И. В., Федорченко Е. В., Красов А. В., Ушаков И. А.** Определение психического состояния пользователей социальной сети Reddit на основе методов машинного обучения. *Информационно-управляющие системы*, 2022, № 1, с. 8–18. doi:10.31799/1684-8853-2022-1-8-18
6. **Браницкий А. А., Дойникова Е. В., Котенко И. В.** Использование нейросетей для прогнозирования подверженности пользователей социальных сетей деструктивным воздействиям. *Информационно-управляющие системы*, 2020, № 1, с. 24–33. doi:10.31799/1684-8853-2020-1-24-33

7. **Stankevich M., Ignatiev N., Smirnov I.** Predicting depression with social media images. *Proc. of the 9th Intern. Conf. on Pattern Recognition Applications and Methods (ICPRAM 2020)*, 2020, pp. 235–240. doi:10.5220/0009168602350240
8. **Stankevich M., Smirnov I., Kiselnikova N., Ushakova A.** Depression Detection from Social Media Profiles. *Data Analytics and Management in Data Intensive Domains. DAMDID/RCDL 2019. Communications in Computer and Information Science*, 2019, no. 1223, pp. 181–194. doi:10.1007/978-3-030-51913-1_12
9. **Pampouchidou A., Simos P. G., Marias K., Meriaudeau F., Yang F., Pediaditis M., Tsiknakis M.** Automatic assessment of depression based on visual cues: A systematic review. *IEEE Transactions on Affective Computing*, 2019, vol. 10, no. 4, pp. 445–470. doi:10.1109/TAFFC.2017.2724035
10. **Wu P., Wang R., Lin H., Zhang F., Tu J., Sun M.** Automatic depression recognition by intelligent speech signal processing: A systematic survey. *CAAI Trans. Intell. Technol*, 2023, no. 8(3), pp. 701–711. doi:10.1049/cit2.12113
11. **Величко А. Н., Карпов А. А.** Аналитический обзор систем автоматического определения депрессии по речи. *Информатика и автоматизация*, 2021, т. 20, № 3, с. 497–529. doi:10.15622/ia.2021.3.1
12. **Ringeval F., Schuller B., Valstar M., Cummins N., Cowie R., Tavabi L., Schmitt M., Alisamir S., Amiriparian S., Messner E.-M., Song S., Liu S., Zhao Z., Malloi-Ragolta A., Ren Z., Soleymani M., Pantic M.** AVEC 2019 Workshop and Challenge: State-of-mind, detecting depression with AI, and cross-cultural affect recognition. *Proc. of the 9th ACM Intern. Workshop on Audio/visual Emotion Challenge (AVEC '19)*, Nice, France, 2019, pp. 3–12. doi:10.1145/3347320.3357688
13. **Sadeghi M., Egger B., Agahi R., Richer R., Capito K., Rupp L. H., Schindler-Gmelch L., Berking M., Eskofier B. M.** Exploring the capabilities of a language model-only approach for depression detection in text data. *Proc. of 2023 IEEE EMBS Intern. Conf. on Biomedical and Health Informatics (BHI)*, Pittsburgh, USA, 2023, pp. 1–5. doi:10.1109/BHI58575.2023.10313367
14. **Величко А. Н.** *Методы и программная система интегрального анализа деструктивных паралингвистических явлений в разговорной речи*: дис. … канд. техн. наук. СПб ФИЦ РАН, 2023. 136 с.
15. **Величко А. Н., Карпов А. А.** Методика и программная система интегрального анализа деструктивных паралингвистических явлений в разговорной речи. *Информационно-управляющие системы*, 2023, № 4, с. 2–11. doi:10.31799/684-8853-2023-4-2-11, EDN: FHUUWJ
16. **Gratch J., Artstein R., Lucas G. M., Stratou G., Scherer S., Nazarian A., Wood R., Boberg J., DeVault D., Marsella S., Traum D., Rizzo S., Morency L.-P.** The distress analysis interview corpus of human and computer interviews. *Proc. of the Ninth Intern. Conf. on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, 2014, pp. 3123–3128.
17. **Ravi V., Wang J., Flint J., Alwan A.** Fraug: A frame rate based data augmentation method for depression detection from speech signals. *Proc. of ICASSP 2022–2022 IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, 2022, pp. 6267–6271. doi:10.1109/ICASSP43922.2022.9746307
18. **Bailey A., Plumbley M. D.** Gender bias in depression detection using audio features. *Proc. of 2021 29th European Signal Processing Conf. (EUSIPCO)*, Dublin, Ireland, 2021, pp. 596–600. doi:10.23919/EUSIPCO54536.2021.9615933
19. **Radford A., Kim J. W., Xu T., Brockman G., McLeavey C., Sutskever I.** Robust speech recognition via large-scale weak supervision. *Proc. of the 40th Intern. Conf. on Machine Learning (ICML'23)*, Honolulu, Hawaii, 2023, no. 202, pp. 28492–28518. doi:10.5555/3618408.3619590
20. **Bredin H.** Pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. *Proc. of INTERSPEECH-2023*, Dublin, Ireland, 2023, pp. 1983–1987. doi:10.21437/Interspeech.2023-105
21. **Boersma P.** Praat, a system for doing phonetics by computer. *Glot International*, 2001, no. 5(9/10), pp. 341–345.
22. **Eyben F., Scherer K. R., Schuller B. W., Sundberg J., Andre E., Busso C., Devillers L. Y., Epps J., Laukka P., Narayanan S. S., Truong K. P.** The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 2015, no. 7(2), pp. 190–202. doi:10.1109/TAFFC.2015.2457417
23. **Huang G., Liu Z., Van Der Maaten L., Weinberger K. Q.** Densely connected convolutional networks. *Proc. of 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, Hawaii, 2017, pp. 2261–2269. doi:10.1109/CVPR.2017.243
24. **Arik S. O., Pfister T.** TabNet: Attentive interpretable tabular learning. *Proc. of the AAAI Conf. on Artificial Intelligence*, Virtual Event, 2021, no. 35(8), pp. 6679–6687. doi:10.1609/aaai.v35i8.16826
25. **Ray A., Kumar S., Reddy R., Mukherjee P., Garg R.** Multi-level attention network using text, audio and video for depression prediction. *Proc. of the 9th ACM Intern. Workshop on Audio/visual Emotion Challenge (AVEC '19)*, Nice, France, 2019, pp. 81–88. doi:10.1145/3347320.3357697
26. **Makiuchi M. R., Warnita T., Uto K., Shinoda K.** Multimodal fusion of BERT-CNN and Gated CNN representations for depression detection. *Proc. of the 9th ACM Intern. Workshop on Audio/visual Emotion Challenge (AVEC '19)*, Nice, France, 2019, pp. 55–63. doi:10.1145/3347320.3357694
27. **Velichko A., Markitantov M., Kaya H., Karpov A.** Complex paralinguistic analysis of speech: Predicting

gender, emotions and deception in a hierarchical framework. *Proc. of INTERSPEECH-2022*, Incheon, Korea, 2022, pp. 4735–4739. doi:10.21437/Interspeech.2022-11294

28. Ушаков И. Б., Бубеев Ю. А., Сыркин Л. Д., Карпов А. А., Поляков А. В., Иванов А. В., Усов В. М. Дистанционное телеконсультирование в первичном звене здравоохранения для скрининга тревожно-депрессивных расстройств с контуром обратной связи от пациента. *Системный анализ и управление в биомедицинских системах*, 2023, № 22(4), с. 140–153. doi:10.36622/VSTU.2023.22.4.022

UDC 004.934.2
doi:10.31799/1684-8853-2024-4-2-11
EDN: RBUXLJ

An approach to depression detection in speech using a semi-automatic data annotation

A. N. Velichko[a], PhD, Tech., Senior Researcher, orcid.org/0000-0002-8503-8512
A. A. Karpov[a], Dr. Sc., Tech., Professor, orcid.org/0000-0003-3424-652X, karpov@iias.spb.su
[a]St. Petersburg Federal Research Center of the RAS, 39, 14th Line, 199178, Saint-Petersburg, Russian Federation

Introduction: Automatic speech recognition tools have been significantly improved since the release of one of the biggest corpora containing speech of the people diagnosed with depression, Extended Distress Analysis Interview Corpus. New methods make it possible to enhance the annotation quality as well as the depression detection quality. **Purpose:** To improve the depression detection quality by combining an automatic audio transcription with getting timestamps for each phrase and manual data validation for error correction if needed. **Results:** We present a semi-automatic approach for data annotation using the Faster-Whisper models for transcribing, a set of scripts for data preprocessing and Praat software for manual data validation. In the experiments we use several different machine learning techniques for classification and regression tasks. An attempt of data normalizing results in enhancing the values for k-Nearest Neighbours method on the preprocessed data but it doesn't involve any significant changes and even worsens the values for the original data. The analysis of the experimental results shows that despite the decrease in the average values, the gap between the values for each class has been reduced. This indicates that the purpose of the work is achieved. **Practical relevance:** We have improved the quality of annotation as well as depression detection in experiments using the presented approach. **Discussion:** In subsequent studies we are going to use this annotation for building the automatic multimodal method for depression detection.

Keywords — speech analysis, speech technologies, computational paralinguistics, destructive phenomena, data preprocessing, automatic depression detection in speech.

For citation: Velichko A. N., Karpov A. A. An approach to depression detection in speech using a semi-automatic data annotation. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2024, no. 4, pp. 2–11 (In Russian). doi:10.31799/1684-8853-2024-4-2-11, EDN: RBUXLJ

Financial support
This work was financially supported by the Russian Science Foundation (project No. 22-11-00321, https://www.rscf.ru/project/22-11-00321/).

References

1. *Depressive disorder WHO (depression).* WHO. 2023. Available at: https://www.who.int/news-room/factsheets/detail/depression (accessed 12 June 2024).
2. Singer K. Depressive disorders from a transcultural perspective. *Social Science & Medicine*, 1975, no. 9, pp. 289–301. doi:10.1016/0037-7856(75)90001-3
3. Alghowinem Sh., Goecke R., Epps J., Wagner M., Cohn J. Cross-cultural depression recognition from vocal biomarkers. *Proc. of INTERSPEECH-2016*, San Francisco, USA, 2016, pp. 1943–1947. doi:10.21437/Interspeech.2016-1339
4. Kaya H., Fedotov D., Dresvyanskiy D., Doyran M., Mamontov D., Markitantov M., Salah A., Kavcar E., Karpov A., Salah A. Predicting depression and emotions in the crossroads of cultures, paralinguistics, and non-linguistics. *Proc. of the 9th ACM Intern. Workshop on Audio/visual Emotion Challenge (AVEC '19)*, Nice, France, 2019, pp. 27–35. doi:10.1145/3347320.3357691
5. Branitskiy A. A., Sharma Y. D., Kotenko I. V., Fedorchenko E. V., Krasov A. V., Ushakov I. A. Determination of the mental state of users of the social network Reddit based on machine learning methods. *Informatsionno-upravliaiushchie sistemy* [*Information and Control Systems*], 2022, no. 1, pp. 8–18 (In Russian). doi:10.31799/1684-8853-2022-1-8-18
6. Branitskiy A. A., Doynikova E. V., Kotenko I. V. Use of neural networks for forecasting of the exposure of social network users to destructive impacts. *Informatsionno-upravliaiushchie sistemy* [*Information and Control Systems*], 2020, no. 1, pp. 24–33 (In Russian). doi:10.31799/1684-8853-2020-1-24-33
7. Stankevich M., Ignatiev N. Smirnov I. Predicting depression with social media images. *Proc. of the 9th Intern. Conf. on Pattern Recognition Applications and Methods (ICPRAM 2020)*, 2020, pp. 235–240.
8. Stankevich M., Smirnov I., Kiselnikova N., Ushakova A. *Depression Detection from Social Media Profiles*. In: *Data Analytics and Management in Data Intensive Domains. DAMDID/RCDL 2019. Communications in Computer and Information Science*, 2019, no. 1223, pp. 181–194.
9. Pampouchidou A., Simos P. G., Marias K., Meriaudeau F., Yang F., Pediaditis M., Tsiknakis M. Automatic assessment of depression based on visual cues: A systematic review. *IEEE Transactions on Affective Computing*, 2019, vol. 10, no. 4, pp. 445–470. doi:10.1109/TAFFC.2017.2724035
10. Wu P., Wang R., Lin H., Zhang F., Tu J., Sun M. Automatic depression recognition by intelligent speech signal processing: A systematic survey. *CAAI Trans. Intell. Technol*, 2023, no. 8(3), pp. 701–711. doi:10.1049/cit2.12113
11. Velichko A., Karpov A. Analytical survey of automatic systems for depression detection by speech. *Informatics and Automation*, 2021, vol. 20, no. 3, pp. 497–529 (In Russian). doi:10.15622/ia.2021.3.1
12. Ringeval F., Schuller B., Valstar M., Cummins N., Cowie R., Tavabi L., Schmitt M., Alisamir S., Amiriparian S., Messner E.-M., Song S., Liu S., Zhao Z., Malloi-Ragolta A., Ren Z., Soleymani M., Pantic M. AVEC 2019 Workshop and Challenge: State-of-mind, detecting depression with AI, and cross-cultural affect recognition. *Proc. of the 9th ACM Intern. Work-*

shop on Audio/visual Emotion Challenge (AVEC '19), Nice, France, 2019, pp. 3–12. doi:10.1145/3347320.3357688
13. Sadeghi M., Egger B., Agahi R., Richer R., Capito K., Rupp L. H., Schindler-Gmelch L., Berking M., Eskofier B. M. Exploring the capabilities of a language model-only approach for depression detection in text data. *Proc. of 2023 IEEE EMBS Intern. Conf. on Biomedical and Health Informatics (BHI)*, Pittsburgh, USA, 2023, pp. 1–5. doi:10.1109/BHI58575.2023.10313367
14. Velichko A. N. *Metody i programmnaja sistema integralnogo analiza destruktivnyh paralingvisticheskih javlenij v razgovornoj rechi*. Dis. kand. tekhn. nauk [Methods and software for integral analysis of destructive paralinguistic phenomena in colloquial speech. PhD Tech. sci. diss.]. Saint-Petersburg, SPb FIC RAN, 2023. 136 p. (In Russian).
15. Velichko A. N., Karpov A. A. An approach and software system for integral analysis of destructive paralinguistic phenomena in colloquial speech. *Informatsionno-upravliaiushchie sistemy* [*Information and Control Systems*], 2023, no. 4, pp. 2–11 (In Russian). doi:10.31799/684-8853-2023-4-2-11, EDN: FHUUWJ
16. Gratch J., Artstein R., Lucas G. M., Stratou G., Scherer S., Nazarian A., Wood R., Boberg J., DeVault D., Marsella S., Traum D., Rizzo S., Morency L.-P. The distress analysis interview corpus of human and computer interviews. *Proc. of the Ninth Intern. Conf. on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, 2014, pp. 3123–3128.
17. Ravi V., Wang J., Flint J., Alwan A. Fraug: A frame rate based data augmentation method for depression detection from speech signals. *Proc. of ICASSP 2022–2022 IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, 2022, pp. 6267–6271. doi:10.1109/ICASSP43922.2022.9746307
18. Bailey A., Plumbley M. D. Gender bias in depression detection using audio features. *Proc. of 2021 29th European Signal Processing Conf. (EUSIPCO)*, Dublin, Ireland, 2021, pp. 596–600. doi:10.23919/EUSIPCO54536.2021.9615933
19. Radford A., Kim J. W., Xu T., Brockman G., McLeavey C., Sutskever I. Robust speech recognition via large-scale weak supervision. *Proc. of the 40th Intern. Conf. on Machine Learning (ICML'23)*, Honolulu, Hawaii, 2023, no. 202, pp. 28492–28518. doi:10.555/3618408.3619590
20. Bredin H. Pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. *Proc. of INTERSPEECH-2023*, Dublin, Ireland, 2023, pp. 1983–1987. doi:10.21437/Interspeech.2023-105
21. Boersma P. Praat, a system for doing phonetics by computer. *Glot International*, 2001, no. 5(9/10), pp. 341–345.
22. Eyben F., Scherer K. R., Schuller B. W., Sundberg J., Andre E., Busso C., Devillers L. Y., Epps J., Laukka P., Narayanan S. S., Truong K. P. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 2015, no. 7(2), pp. 190–202. doi:10.1109/TAFFC.2015.2457417
23. Huang G., Liu Z., Van Der Maaten L., Weinberger K. Q. Densely connected convolutional networks. *Proc. of 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, Hawaii, 2017, pp. 2261–2269. doi:10.1109/CVPR.2017.243
24. Arik S. O., Pfister T. TabNet: Attentive interpretable tabular learning. *Proc. of the AAAI Conf. on Artificial Intelligence*, Virtual Event, 2019, vol. 35, no. 8, pp. 6679–6687. doi:10.1609/aaai.v35i8.16826
25. Ray A., Kumar S., Reddy R., Mukherjee P., Garg R. Multi-level attention network using text, audio and video for depression prediction. *Proc. of the 9th ACM Intern. Workshop on Audio/visual Emotion Challenge (AVEC '19)*, Nice, France, 2019, pp. 81–88. doi:10.1145/3347320.3357697
26. Makiuchi M. R., Warnita T., Uto K., Shinoda K. Multimodal fusion of BERT-CNN and Gated CNN representations for depression detection. *Proc. of the 9th ACM Intern. Workshop on Audio/visual Emotion Challenge (AVEC '19)*, Nice, France, 2019, pp. 55–63. doi:10.1145/3347320.3357694
27. Velichko A., Markitantov M., Kaya H., Karpov A. Complex paralinguistic analysis of speech: Predicting gender, emotions and deception in a hierarchical framework. *Proc. of INTERSPEECH-2022*, Incheon, Korea, 2022, pp. 4735–4739. doi:10.21437/Interspeech.2022-11294
28. Ushakov I. B., Bubeev Yu. A., Syrkin L. D., Karpov A. A., Polyakov A. V., Ivanov A. V., Usov V. M. Remote tele-counseling in primary healthcare for screening of anxiety-depressive disorders with a feedback loop from the patient. *System Analysis and Management in Biomedical Systems*, 2023, vol. 22, no. 4, pp. 140–153 (In Russian). doi:10.36622/VSTU.2023.22.4.022