Вопросы лексикографии. 2025. № 36. С. 67–103
Russian Journal of Lexicography. 2025, 36, pp. 67–103

Научная статья
УДК 81’322.2
doi: 10.17223/22274200/36/4

Датасет как форма словаря в цифровую эпоху (на примере мультимодального эмоционального датасета)

Анастасия Владимировна Колмогорова[1], Елизавета Романовна Куликова[2]

[1,2] Национальный исследовательский университет «Высшая школа экономики», Санкт-Петербург, Россия
[1] akolmogorova@hse.ru
[2] Kulikova.E.R@hse.ru

Аннотация. Статья посвящена аргументации тезиса о том, что в современном технологичном обществе у традиционного лингвистического словаря появился новый системный вариант – датасет. Имея общую логику, выстроенную по принципу «объект – ключ», словарь и датасет обладают и некоторыми отличиями, которые мы иллюстрируем на примере мультимодального датасета эмоций. Он предназначен для исследования эмоциональной речи на русском языке и оценки качества автоматического детектирования эмоций в различных модальностях с использованием компьютерных моделей. Основная цель статьи – продемонстрировать потенциал датасета как новой формы систематизации и манифестации экспертного знания лингвиста в цифровую эпоху.

Ключевые слова: мультимодальность, датасет, словарь, автоматический анализ эмоций, мультимодальная разметка, речь, русский язык, цифровая лингвистика

Благодарности: Статья подготовлена в ходе проведения исследования в рамках программы фундаментальных исследований Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ).

Для цитирования: Колмогорова А.В., Куликова Е.Р. Датасет как форма словаря в цифровую эпоху (на примере мультимодального эмоционального датасета) // Вопросы лексикографии. 2025. № 36. С. 67–103. doi: 10.17223/22274200/36/4

© Колмогорова А.В., Куликова Е.Р., 2025

Original article
doi: 10.17223/22274200/36/4

Dataset as a form of a dictionary in the digital age (as based on the example of a multimodal emotion dataset)

Anastasia V. Kolmogorova[1], Elizaveta R. Kulikova[2]

[1, 2] Higher School of Economics, Saint Petersburg, Russian Federation
[1] akolmogorova@hse.ru
[2] Kulikova.E.R@hse.ru

Abstract. The article substantiates the thesis that in modern technological society, the traditional linguistic dictionary has acquired a new systemic variant – the dataset. While sharing a common "object–key" principle logic, dictionaries and datasets also possess certain differences, which we illustrate using the example of a multimodal emotion dataset. It is designed for studying emotional speech in Russian and assessing the quality of automatic emotion detection across various modalities using computer models. The article aims to demonstrate the potential of datasets as a new form of systematizing and manifesting linguists' expert knowledge in the digital era. The corpus comprises 173 minutes of video recordings of emotional narratives collected using the autobiographical MIP method with the participation of eleven women aged 19–26. The recorded emotional videos were divided into 909 fragments. Each was annotated on six emotional scales (joy, sadness, anger, surprise, fear, disgust) on a 0–5 scale by six annotators (three annotators worked with one half of the sample, three with the other half) in four formats: multimodal and separate audio, text, and video fragments. The key findings from the dataset analysis are as follows. (1) When comparing modalities, the highest inter-annotator agreement scores were observed for text annotations and full multimodal annotations ($\alpha = 0.57$), while the lowest was for video-only annotations ($\alpha = 0.30$). (2) When comparing annotator consistency metrics across emotional classes, the highest agreement was found in assessing neutral texts; agreement was relatively high for joyful and sad texts, while mixed emotions were recognized least consistently. (3) Joy and surprise are primarily recognized when fragments are presented in audio format; sadness, fear and disgust are better identified in audio and text modalities, while anger is most accurately recognized only in text modality. (4) Presenting fragments in video format reduces recognition accuracy for all emotions, with the least impact on joy and the greatest on fear. The dataset has also proven effective as a tool for evaluating eight computer emotion recogni-

tion models, including text, audio and multimodal models. The highest alignment with human annotations was shown by text-based models, while the worst results came from video-based models. Despite some limitations related to data collection and speech segmentation, the dataset represents a valuable linguistic resource for emotion recognition research.

Keywords: multimodality, dataset, dictionary, automatic emotion analysis, multimodal markup, speech, Russian language, digital linguistics

Acknowledgments: The article was prepared within the framework of the Basic Research Program at HSE University.

For citation: Kolmogorova, A.V. & Kulikova, E.R. (2025) Dataset as a form of a dictionary in the digital age (as based on the example of a multimodal emotion dataset). *Voprosy leksikografii – Russian Journal of Lexicography*. 36. pp. 67–103. (In Russian). doi: 10.17223/22274200/36/4

Введение

Создание словарей – важная прикладная задача лингвистики. Лексикография имеет многовековую традицию: толковые, этимологические, орфоэпические и иные словари являются авторитетным источником эталонного лингвистического знания, к которому носители языка обращаются за «подсказками» в случае проблем при интерпретации словоупотребления, встреченного ими в текстах.

Однако новое, цифровое, общество активно формирует привычку к использованию компьютерных инструментов для принятия того или иного интерпретативного решения: если у носителя языка есть сомнения относительно смысла текста, у него есть также и выбор – обратиться к словарю или отправить запрос Большой языковой модели (далее – БЯМ). Однако чтобы БЯМ или иная нейросеть справилась с задачей, ей не нужна отвлеченная от текстовой среды абстрактная информация о слове – ей необходимы валидные примеры проявления некоторого искомого феномена непосредственно в тексте.

В данной публикации мы ставим своей целью системно представить результаты, полученные нами в ходе экспериментальной работы по разметке эмоций в речевом материале на русском языке, очертив тем самым потенциал использования собранных и систематизированных данных в исследованиях эмоциональной речи.

Датасет как новая форма лингвистического словаря

Лингвистическое описание эмоций имеет богатую историю в отечественной науке.

С легкой руки В.И. Шаховского в конце XX в. в языкознании появилось новое направление исследований – лингвоэмотиология. Постулировав, что весь язык эмоционален [1], исследователь определил в качестве основной задачи лингвоэмотиологии изучение того, как эмоциональное состояние человека проявляется и преломляется в речи и шире – коммуникативном поведении носителя определенного языка. Таким образом, было сформулировано априори междисциплинарное кредо лингвоэмотиологии – изучить языковую категорию эмотивности в тесной связи с внутренним миром человека.

Еще более широкое проблемное поле все в том же междисциплинарном русле очертила для себя лингвопсихология – изучать «психологию человека, его внутренний мир – эмоции и чувства, поведение, качества личности, выявляемые и интерпретируемые с позиций лингвистики и с использованием ее методологии» [2, с. 267]. В русле данного направления, например, выполнен словарь-тезаурус эмотивной лексики [3], содержащий функциональные и дискурсивные интерпретации 39 денотативно-идеографических групп эмотивной лексики русского языка. В частности, Л.Г. Бабенко показала на обширном материале, что эмоции могут интерпретироваться человеком и концептуализироваться в языке через призму целого ряда категориальных рамок: состояние (грусть), отношение (любить), воздействие на кого-либо (влюбить), становление эмоции (влюбиться, разлюбить), характеризация (влюбчивый), качество (влюбчивость), человек – носитель эмоций (любимчик, любовник, любитель) [3, с. 274]. С данной структурой согласуется и последовательность рубрик в словарной статье – после обобщенного описания ситуации переживания эмоции, ее объекта, интенсивности и длительности следуют перечни лексем разных грамматических классов, которые в своей семантике в той или иной мере отражают представленные выше категориальные рамки концептуализации эмоции (состояние, становление, характеризация и т. д.). Данный словарь – замечательный пример лексикографирования эмоциональной лексики русского языка.

Датасет – это файл, размер и формат которого может варьироваться, в нем системно представлена информация, структурированная с помощью колонок и строк. В отличие от базы данных датасет одномерен и статичен – он не дает возможности управлять данными (извлекать информацию, запрашивать ее посредством обращения к категориям, переструктурировать). В определенном смысле датасет – это также вид словаря, где представлены пары «объект – ключ». Например, в классическом толковом словаре объекту соответствует лемма, ключу – ее толкование. В датасете же объектом становится текстовый или речевой фрагмент, ключом – его категориальная метка. Например, датасет, посвященный коммуникативному феномену токсичности, может содержать примеры текстов (объект), каждый из которых будет иметь одну из следующих меток (ключ): нетоксичный, токсичный, очень токсичный. Таким образом, мы не даем готового определения токсичности, но демонстрируем примеры текстов, в которых по оценкам аннотаторов данный феномен представлен в той или иной степени либо отсутствует. В этом смысле датасет возвращает нас, хотя и на новом уровне, к истокам лексикографии.

Первыми лексикографическими практиками были практики глоссирования церковных текстов, когда на полях рукописей составлялись краткие комментарии к непонятному слову или сложному обороту. Позже списки пар «слово языка 1 – его эквивалент на языке 2» или «слово – его толкование» стали составляться отдельно от тех текстов, в которых они встречались. Такое «отвлечение слов от непосредственной связи с текстом придало им более обобщенный характер» [4, с. 41], ознаменовав собой собственно начало словарного дела. Своеобразное изъятие слова из родной среды бытования – речи и текста – было совершенно естественно в ситуации, когда интерпретатором слова в некотором его употреблении выступал человек, рядовой носитель языка. У него не было возможности прочитать большие объемы текстовых данных, но нужны были «ключи» понимания, готовые к употреблению, – словарные дефиниции. В то же время создание таких пар «объект – ключ» для всеобщего использования стало профессиональной задачей специалистов-лексикографов – именно им доверено читать относительно большие объемы текстов, «отчуждать»

от них конкретное слово или слова, анализировать и обобщать, чтобы сформулировать, что значит данное слово «вообще».

Однако на том этапе развития технологий, на котором сегодня находится лингвистика, в частности – компьютерная, появились возможности в сжатые сроки обрабатывать очень большой объем текстов. Следовательно, рядовому носителю языка в ряде случаев уже не нужно прибегать к посредничеству словаря, чтобы интерпретировать то или иное словоупотребление в тексте, он может делегировать эту функцию, скажем, БЯМ. В такой ситуации модифицируется и профессиональная задача лингвиста – сегодня он востребован как компетентный «поставщик» релевантных данных для того, чтобы обучить модель интерпретации. Ему нужно найти и систематизировать текстовые/речевые образцы, проверить их при помощи экспертной оценки, выставить метки и предложить их модели. Она обнаружит некоторый паттерн, не отвлеченный от текста, а глубоко фундированный в нем, благодаря которому типичный носитель языка приписывает данному тексту определенный смысл, и будет аналогичным образом применять его на других текстах.

Иными словами, используя датасет, мы вновь обращаемся к идее своеобразного глоссирования текстов, возвращая конкретные слова в лоно текстовой среды.

Так, в собранном нами датасете в качестве объектов выступают текстовые/речевые/жестово-мимические и мультимодальные единства небольшого объема, в качестве ключей – числовые эмоциональные значения, агрегированные из множества значений, выставленных каждому из данных объектов аннотаторами. Если, например, в словаре «Алфавит эмоций» Л.Г. Бабенко мы можем непосредственно посмотреть, указание на какую эмоцию в том или ином объеме содержит в себе семантика тех или иных лексических единиц русского языка по данным их лексикографических описаний, то из нашего датасета мы получаем информацию о том, какое эмоциональное значение (для каждой эмоции – в числовом выражении в интервале от 0 до 5) приписывается некоторому ограниченному, но когерентному множеству контекстуально связанных лексических единиц, предъявляемых в разных модальностях. Недостатком такой организации является невозможность напрямую получить ответ на вопрос, какой эмо-

циональный вес имеет каждая конкретная лексема. Достоинством же считаем то, что в данном случае преодолевается проблема дискретности значений лексических компонентов – используя датасет, мы можем дообучить компьютерные нейросетевые модели для «узнавания» эмоций в других подобных коллекциях речевых данных на русском языке в разных модальностях (текстовой, звучащей речи или видеозаписи) или использовать в качестве «золотого стандарта» для оценки качества автоматического анализа эмоций в таких речевых данных с использованием существующих моделей.

Проиллюстрируем нашу мысль о преодолении дискретности значений при помощи датасета следующим образом. Ниже представлен текст поста в одной из социальных сетей:

«Мне 25 лет. Уже год живу с молодым человеком раздельно и за это время я привыкла жить одна: не жду его в гости, не хочу спать с ним и впускать в свою зону комфорта. Я больше не хочу секса, поцелуев и объятий. Я хочу быть одной».

Если мы зададимся вопросом, какое слово в данном тексте передает в наибольшей степени смысл ‘грустить, тосковать’, и последовательно переберем в словаре толкования каждой из составляющих данный текст лемм, то не найдем ни одной, которая бы содержала данный смысловой компонент на уровне лексического значения. Однако разметчики, оценивая фрагмент, уверенно маркируют его как грустный. Очевидно, эмоциональная тональность грусти манифестируется не столько в семантике отдельного слова, сколько через синтагматику форм: повторы глаголов с отрицательной частицей, при этом отрицанию подвергаются аргументы, имеющие положительную коннотацию (не хочу поцелуев и объятий); повторы местоименной формы одна; перечисления и параллельные конструкции. Иными словами, «отвлечение» значения слова от текстовой среды в данном случае оказывается малоэффективным, а анализ категоризированной (мы знаем благодаря агрегированному мнению разметчиков, что это грустный текст) текстовой среды, напротив, – более продуктивным.

Мультимодальный же датасет позволяет нам выходить за пределы собственно языковых/речевых средств манифестации какого-либо смысла, включая в контур интерпретации паралингвистические факторы.

Существующие мультимодальные датасеты: тематический обзор

Мультимодальность – термин, достаточно недавно вошедший в исследовательскую практику преимущественно гуманитарных наук. В лингвистике и семиотике все разнообразие его трактовок можно свести к трем основным точкам зрения: мультимодальность – это восприятие человеком некого внешнего стимула одновременно посредством зрения и слуха [5]; мультимодальность – «совмещение более одного способа репрезентации объектов семиотическими средствами и/или использование нескольких каналов (аудиального, визуального и проч.)» [6, с. 23]; наконец, мультимодальность – это устный и письменный язык плюс все остальное, что языком не является, но также служит семиотическим средством выражения смыслов в коммуникации [7].

Исследователи же, работающие в сфере автоматического распознавания эмоций, машинного обучения, а также корпусной лингвистики, активно обсуждают проблемы мультимодальных корпусов [8], мультимодальных систем и моделей [9], подразумевая под мультимодальностью специфическую характеристику используемых данных – такие данные объединяют аудио, видео и текст. В англоязычных работах в тех случаях, когда имеется в виду вся совокупность информации, которая может быть автоматически извлечена из акустических, визуальных и текстовых данных, говорят о звуковой, визуальной и текстовой модальностях (audio, visual, and textual modalities) [10].

Таким образом мультимодальным можно назвать такой лингвистический датасет, который содержит по меньшей мере два типа данных из перечисленных: аудио, тексты (например, в форме текстовой расшифровки записи), видеозапись речевого поведения, запись движений по технологии motion capture, данные о физиологических показателях в момент речепроизнесения (пульс, давление, сердечный ритм и т. д.).

Существующие мультимодальные датасеты отличаются друг от друга по следующим критериям:

1) тип получения эмоциональных данных: симуляции, выполненные актерами в лабораторных условиях [11]; фрагменты медиаконтента (фильмы, сериалы, телевизионные передачи, записи на YouTube, ток-шоу) [12–14]; экспериментальные процедуры, предпо-

лагающие индуцирование эмоционального состояния у информанта посредством демонстрации эмоциональных стимулов [15] (изображений, видеофрагментов, текстов) или посредством методов, основанных на психологических механизмах эмпатии и интроспекции [16]; записи спонтанной коммуникации в естественных условиях [17];

2) типология эмоций, используемая для дизайна разметки. Если исследователи опираются на категориальный подход, то рассматривают сферу эмоций как пространство, поделенное на конечное число категорий с четкими границами – грусть, радость и т. д. Число категорий варьируется от 6 в монографии П. Экмана и Р. Дэвидсона [18] или 8 – в ряде других работ [19–21] до 11 в недавно опубликованной статье Wang et al. [22]. Если для типологии эмоций используют пространственный подход, то трактуют каждую эмоцию как некоторую область в эмоциональном континууме, координаты которой задаются тремя измерениями, соответствующими неким параметрам, не являющимся эмоциями, но предопределяющими их (например, VAD-модель, в которой эмоциональная область задается параметрами valence (оценка эмоционального состояния), arousal (возбуждение), dominance (социальное доминирование)). Так, например, в [22] используется датасет на основе категориального подхода – разметчики оценивают присутствие 6 базовых эмоций в видеомонологах, выбирая одну или несколько эмоций из предложенного списка. В [23] применяется пространственный подход, при котором разметчикам не предоставляют никакого списка категорий – им предлагается оценить объекты по следующим шкалам (от 0 до 7): согласие между участниками эмоционального диалога, их стремление к доминированию, вовлеченность во взаимодействие, эффективность коммуникации и способность к установлению межличностного контакта. Таким образом, при пространственном подходе объекты эмоционального аннотирования не помещаются в рамки жестко заданных категорий, а характеризуются относительно своего места в многомерном пространстве, задаваемом шкалами;

3) методика агрегации результатов разметки. Поскольку восприятие эмоций крайне субъективно, во всех проанализированных нами датасетах исследователи прибегают к аннотации с перекрестным покрытием: один и тот же фрагмент оценивается несколькими размет-

чиками (обычно тремя, иногда больше), после чего рассчитывается итоговая эмоциональная оценка фрагмента. Методы подсчетов итогового эмоционального «значения» объекта наблюдения варьируют в зависимости от подхода. В случае категориального подхода чаще всего исследователи опираются на принцип большинства (majority voting). Если применяется пространственная модель, постобработка может включать использование z-score нормализации по формуле $(x - \mu)/\sigma$, где x – исходное значение, μ – среднее значение данных, σ – стандартное отклонение данных [24]. Чтобы получить ground truth label (эталонные данные) для фрагмента, вычисляются средние значения оценок разметчиков по каждой из шкал.

Проведенный анализ существующих датасетов позволил выявить несколько значимых лакун.

Во-первых, подавляющее их большинство выполнено на материале английского языка – в нашей выборке оказалась только две коллекции данных на русском языке: корпус RAMAS (Russian Multimodal Corpus of Dyadic Interaction for studying emotion recognition) и корпус видеофрагментов, собранный А. Котовым и его группой [17].

Во-вторых, во всех проанализированных нами датасетах разметка осуществляется гомогенно, т. е. разметке подвергается полностью видеофрагмент определенной длительности, а присваиваемая метка автоматически распространяется на все модальности, т. е. считается, что эмоция X во фрагменте N считывается на основе информации, которую мы получаем как из акустической, так и визуальной модальности. Однако известно, что разные каналы (просодический, жестовый, мимический, вербальный) вносят неодинаковый вклад в формирование и, соответственно, восприятие эмоции: «При разработке интеллектуальных систем крайне важно задействовать разные модальности, поскольку некоторые эмоции могут лучше идентифицироваться с опорой на конкретную модальность. Например, уровень эмоционального возбуждения может лучше распознаваться через речевую просодию, в то время как валентность через вербальный канал и визуальную информацию» [25, с. 388]. Иными словами, одна метка фрагмента не отражает особенности взаимодействия разных модальностей в проявлении и узнавании эмоций.

В-третьих, мы не обнаружили прецедентов использования мультимодальных эмоциональных датасетов для оценки качества существующих моделей автоматического анализа эмоций в каждой из модальностей (текстовой, аудиальной (звучащая речь), жестово-мимической), хотя такие модели существуют.

Созданный нами датасет призван заполнить все перечисленные выше лакуны.

Материал и методы. Описание датасета

Создание датасета включает в себя ряд обязательных этапов: 1) проектирование структуры датасета; 2) сбор эмоционального материала; 3) процедура разметки и ее реализация. В сборе материала задействованы носители языка без специального актерского образования (далее – информанты). На третьем этапе для эмоциональной аннотации привлекаются разметчики.

Проектирование структуры датасета

Датасет создавался с целью сформировать такой источник данных об эмоциональной речи, который бы позволил обучать компьютерные модели детектированию эмоций на основе признаков, извлеченных как из одной модальности (канала коммуникации), так и из нескольких. В качестве таргетных каналов коммуникации мы выбрали три: текст, звук (аудиозапись речи), мимика (без аудиосопровождения). В качестве эталонных данных оценивания (ground truth label) мы рассматривали эмоциональные метки, полученные для объекта оценки, представленные разметчику во всех модальностях одновременно (видеофрагмент со звучащей речью и мимикой информанта). Таким образом, в датасете представлены четыре типа объектов оценивания, каждый из которых формирует отдельный сет.

В каждом сете от каждого разметчика в результате аннотирования мы получаем по 8 значений оценки: 6 значений для каждой из базовых эмоций по П. Экману и Р. Дэвидсону (радость, грусть, удивление, отвращение, страх, злость); 1 значение – оценка по шкале нейтральности, еще 1 значение – это оценка по шкале «другое», которую разметчик волен назвать самостоятельно. «Измерением» считался фраг-

мент в одной модальности, оцененный по одной шкале (например, аудиофрагмент по шкале «радость»).

Для каждого измерения имеются оценки 3 аннотаторов. На основе этих оценок подсчитываются и фиксируются в датасете среднее и медианное значения оценки. Также указывается значение альфы Криппендорфа – статистической меры, позволяющей высчитывать процент согласованности мнений разметчиков, если число последних больше двух [26]. В табл. 1 представлен фрагмент структуры датасета для эмоции «радость». Аналогичные подсчёты были сделаны по остальным эмоциям.

Процедура сбора эмоционального материала для разметки

При выборе процедуры сбора эмоционального материала принимались во внимание следующие критерии: 1) баланс между контролируемостью условий и естественностью эмоций, испытываемых информантом; 2) возможность получения развернутой вербальной реакции участника; 3) применимость процедуры индуцирования эмоции

Таблица 1

Фрагмент структуры датасета

ID объекта	Тип модальности	Среднее значение оценки по шкале «радость»	Медианное значение оценки по шкале «радость»	α Криппендорфа для шкалы «радость»
1	1_text (ссылка на папку хранения)	2,3	2	0,3
1	1_audio (ссылка на папку хранения)	3	3	0,4
1	1_video (ссылка на папку хранения)	2,3	2	0,2
1	1_full (ссылка на папку хранения)	3,3	3	0,4

для шести базовых эмоций (радость, грусть, удивление, страх, злость, отвращение).

Мы отказались от применения популярных методов, основанных на сценарном взаимодействии и «разыгрывании» эмоции в заданных ситуациях, поскольку даже при отсутствии прописанных реплик такой экспериментальный сеттинг все же подразумевает компонент актерской игры, имитации, а не переживания эмоции. В связи с этим мы обратились к применяемым в психологии методам индуцирования эмоциональных состояний. Чтобы получить от участника развернутый эмоциональный нарратив, мы не использовали такие техники, как демонстрация в качестве эмоциогенных стимулов изображений, видео или прослушивание музыки, а также чтение текстов про себя. Во время экспериментальных сессий с применением вышеперечисленных методов обнаружилось, что индуцированная такими стимулами эмоция слабо и недостаточно вербализуется, превращаясь в безэмоциональный пересказ событий или в описание переживаний героя видео, которые могут не соотноситься с собственными эмоциями информанта. В итоге для индуцирования эмоциональных состояний была разработана процедура, основанная на автобиографическом методе (Autobiographical MIP) [27].

Информант получал следующую инструкцию: «Подумайте о чем-то, что когда-то вызвало или обычно вызывает у вас... (название эмоции). Постарайтесь погрузиться в эту ситуацию и прочувствовать эмоцию, которую вы в ней испытываете. Вспомните как можно больше деталей. При желании сделайте письменные заметки. Когда будете готовы, расскажите об этой ситуации. Постарайтесь передать ту эмоцию, которую вы в этой ситуации испытали/испытываете».

Такая инструкция повторялась каждому информанту для каждой из 6 эмоций, которые индуцировались в следующем порядке: радость, удивление, злость, страх, грусть и отвращение. При таком подходе мы апеллируем к личному эмоциональному опыту информанта, что увеличивает естественность эмоционального ответа.

В записи датасета приняли участие 11 информантов женского пола, возраст 19–26 лет. Все информанты дали свое письменное согласие на участие. Материал двоих участников был исключен из дальнейшего анализа, так как он был оценен нами как недостаточно ре-

презентативный для исследуемого набора эмоций. Запись проходила в отдельном тихом помещении, информант располагался статично на нейтральном светлом фоне. Кадр был выровнен таким образом, чтобы были видны лицо и плечи информанта, жестикуляция не захватывалась (рис. 1, 2).

Видеозапись осуществлялась на камеру с разрешением 1920 × 1080 и частотой 60 кадров в секунду; для записи звука использовался беспроводной микрофон-петличка. Общий объем корпуса эмоциональных видеозаписей составил 173 мин (2,8 ч).

Рис. 1. Сеттинг эксперимента

Рис. 2. Позиция информанта в кадре

Процедура разметки

Собранные видеозаписи нарративов были разделены на фрагменты-высказывания по двум критериям: 1) завершенная интонация, которая контролировалась по осцилограммам в программной среде ELAN; 2) эмоциональная когерентность фрагмента, которая определялась экспертно, что составляет одно из ограничений исследования, которое мы осознаем. С помощью ELAN вручную были расставлены таймкоды-маркеры границ фрагментов, затем таймкоды экспортировались в текстовом формате для проведения дальнейшей нарезки видеозаписей на фрагменты с использованием библиотеки moviepy (Python). Таким образом, было получено 909 фрагментов. Средняя длина фрагмента – 8,92 с.

Все фрагменты датасета представлены в трех формах: видео + звук (оригинальный фрагмент), только видео (без звука), только аудио, только текст (проверенная экспертно транскрибация[1] высказывания). Для разметки мы выбрали 6 эмоциональных категорий по П. Экману

[1] Под транскрибацией в компьютерной лингвистике понимается автоматическая расшифровка звучащей речи с помощью специальных моделей.

и Р. Дэвидсону: радость, грусть, злость, удивление, отвращение, страх. Кроме того, включена седьмая категория – «нейтрально», а также восьмая – «другое». Поскольку процедура записи эмоциональных нарративов подразумевала высокую степень свободы участников в построении монолога, у информантов возникали эмоции разной силы, а также смешанные эмоции. В связи с этим для каждой эмоциональной категории, в том числе «нейтрально», использовалась шкала от 0 (эмоция не выражена) до 5 (высокая степень выраженности эмоции). Разметчики имели возможность выбрать не одну, а несколько эмоций для фрагмента, чтобы отразить смешанные эмоции.

Для организации разметки использовалась платформа Label Studio [28]. Инструмент имеет версию open source и расширенный вариант. Для проведения академических исследований доступ к расширенной версии предоставляется разработчиками бесплатно по запросу, чем мы и воспользовались.

На этапе реализации интерфейса были решены следующие подзадачи:

1) с использованием синтаксиса XML разметки запрограммированы 4 варианта интерфейса для разных типов аннотируемых данных: текста, аудио, видео (без звука), видео (со звуком);

2) данные выгружены в облачный сервис (Google Cloud);

3) облачный сервис подключен к Label Studio.

Перед началом разметки разметчикам предлагалось ответить на вопросы адаптированной для русского языка «Торонтской шкалы на алекситимию» [29] – в результате 2 разметчика, отнесенные по результатам опросника к группе людей с уровнем алекситимии выше среднего (от 66 баллов и выше), были исключены из выборки разметчиков. При этом мы исходили из предположения, что высокий уровень алекситимии – неспособности личности дифференцировать свои и чужие эмоциональные состояния – снижает валидность разметки, осуществляемой аннотатором.

Всего в разметке приняло участие 6 аннотаторов – женщины с неоконченным или оконченным высшим образованием гуманитарной специальности (средний возраст 21,2 года). Группа разметчиков и фрагменты для аннотации были разделены пополам: три разметчика

работали с первой половиной датасета, еще три – со второй. Таким образом, весь датасет был размечен с тройным покрытием.

Процедура разметки выглядела следующим образом: сначала разметчик аннотирует текстовые фрагменты (расшифровки говоримого в рамках данного фрагмента, полученные автоматически посредством транскрибации, но затем проверенные и скорректированные вручную), затем фрагменты с видеорядом, но без звука, затем – аудиофрагменты звучащей речи, после чего – полные мультимодальные фрагменты. При этом объекты оценки подавались для разметки рандомизированно, чтобы снизить эффект влияния контекста предыдущих объектов оценивания. В начале работы разметчику предъявлялась инструкция (рис. 3).

На рис. 4 представлен интерфейс разметчика в момент разметки эмоции в аудиофрагменте: ему предлагается прослушать аудио, затем расставить, используя слайдер, значения интенсивности эмоций по шкалам от 0 до 5 и нажать на «отправить». После чего перед разметчиком появляется следующий фрагмент для оценки. Объем собранного датасета представлен в табл. 2.

Таким образом, было размечено 3 636 объектов оценки, представленных в разных модальностях.

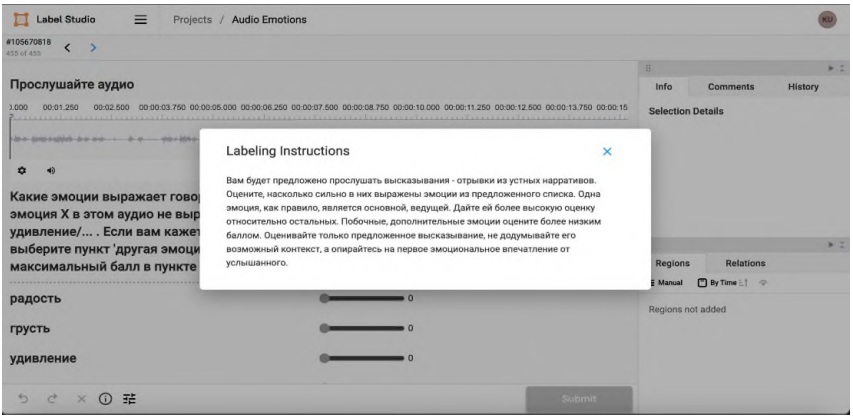

Рис. 3. Пример инструкции при разметке аудио

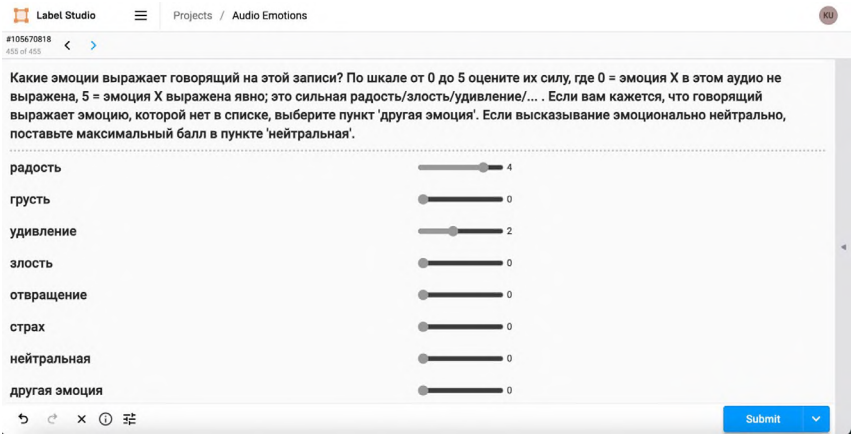

Рис. 4. Интерфейс разметчика для оценки эмоций в аудиофрагменте

Таблица 2

Объем мультимодального эмоционального датасета

Количество/ тип данных	Количество эмоциональных меток	Количество объектов оценки
В каждой из четырех модальностей	43 632	909
Всего во всех четырех модальностях	174 528	3 636

Результаты и обсуждение

Рассмотрим те наблюдения и тренды, которые становятся заметны при анализе датасета.

Согласованность оценок vs модальность

Наибольшие средние значения альфы Криппендорфа были получены при разметке в двух условиях (табл. 3): разметка текстовых

Таблица 3

Средние значения альфы Криппендорфа для фрагментов разных модальностей

Мультимодальный (полный) фрагмент	0,57
Текст	0,57
Аудио	0,46
Видео	0,30

фрагментов и полных фрагментов (α = 0,57), когда разметчики опирались на всю совокупность мультимодальной информации – видео + аудио. Наименьшая согласованность наблюдалась при разметке видеофрагментов без звука – только с опорой на анализ мимической составляющей (0,30).

Проанализируем распределение объектов разметки, поданных в разных модальностях, на шкале значений альфы Криппендорфа (рис. 5): α в промежутке от 0,7 до 1 (максимальное согласие) получили 43,8% мультимодальных фрагментов (на рис. 5 обозначены как

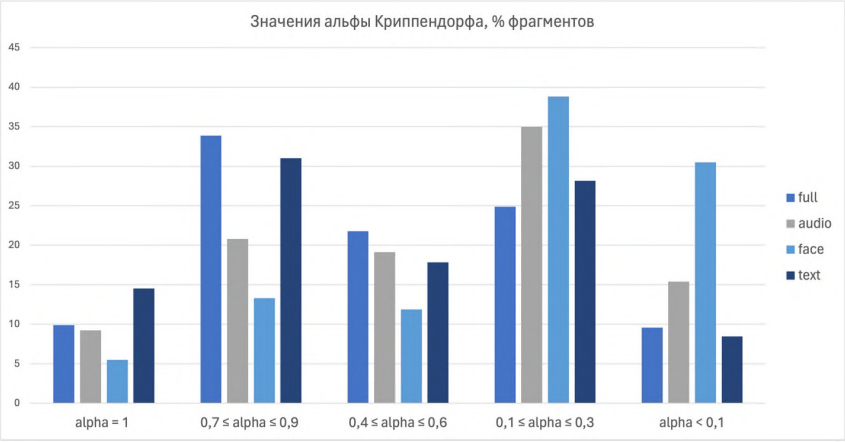

Рис. 5. Распределение значений альфы Криппендорфа

«full»), 45,5% текстовых фрагментов («text»), 30% аудиофрагментов («audio») и 18,8% фрагментов, которые оценивались только по визуальной модальности («face»). Последний тип фрагментов, напротив, превалирует в группе объектов разметки с самой низкой согласованностью (alpha < 0,1).

Согласованность vs эмоции

Наибольшую согласованность демонстрируют эмоционально нейтральные фрагменты. В оценках мультимодальных (полных) фрагментов высокая степень согласия (значение α находится в промежутке от 0,7 до 1) была достигнута для 54,7% радостных и 44,5% грустных фрагментов. Остальные эмоции имеют гораздо более низкую согласованность (<0,5).

Низкие значения альфы (0,2–0,3) получили те фрагменты, где в силу слабой интенсивности эмоции разметчики колебались между нейтральностью и эмоциями грусть/другая/злость/радость. Наиболее низкие значения согласованности (0,1–0,2) имеют фрагменты, в которых разметчики делали выбор между эмоциями близкого спектра: злость/отвращение и злость/грусть.

Обратимся к статистике распределения фрагментов по эмоциям.

Эмоции vs количество объектов оценивания

На рис. 6 представлено распределение эмоций в датасете на основании оценок мультимодальных фрагментов, которые мы рассматривали как ground truth.

Около 30% фрагментов были оценены как эмоционально нейтральные, наиболее сбалансированным оказалось распределение радости (19%), грусти (24%) и злости (18%). Из этого можем сделать вывод, что выбранная нами процедура индуцирования эмоций оказалась наиболее результативной именно для этих трех эмоций. Сложнее оказалось вызвать страх (9%): получая задание вспомнить ситуацию или факт, которые вызывали/вызывают страх, информанты зачастую рассказывали о страхах, которые они имеют в целом, поэтому в процессе монолога их эмоциональное состояние скорее квалифицировалось как тревога или злость.

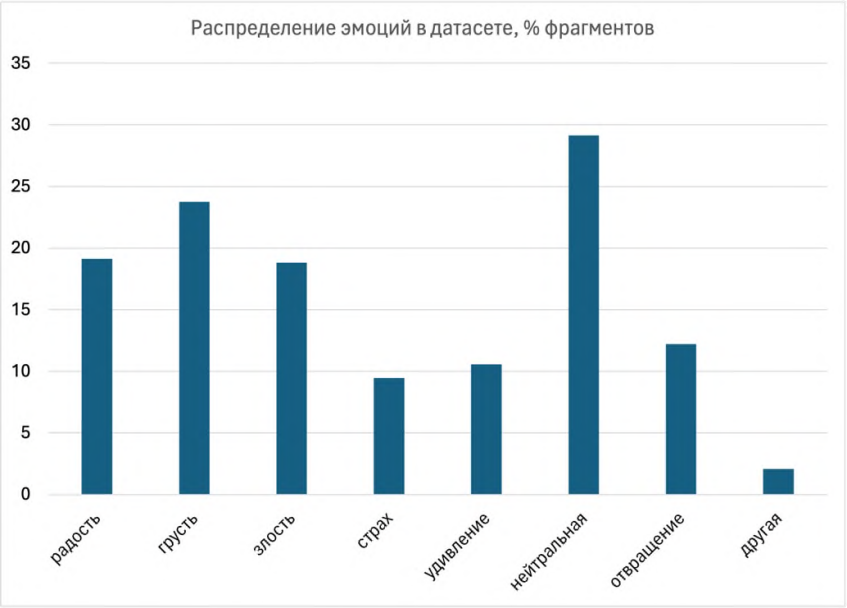

Рис. 6. Распределение эмоций в датасете

Надо сказать, что достаточно большую долю (для мультимодальных фрагментов – 23,7%) составили фрагменты, оцененные как выражающие смешанную эмоцию. В случае опоры только на акустическую модальность – 18,15% фрагментов, только на визуальную – 23,4%, только на текстовую – 25,9% от всех текстовых фрагментов. Учитывая низкую согласованность в оценках по визуальной модальности, высокая доля смешанных фрагментов может быть вызвана расхождением мнений разметчиков: они отмечали разные наборы эмоций, и их усредненное мнение отразилось в итоговой оценке. Для мультимодальных же и текстовых фрагментов согласованность была выше, в связи с чем оценка их как смешанных более надежна.

Например, для текстового фрагмента «Думаю: “Ну как так? Мы с тобой почти пять лет вместе, а ты вот так вот легко сдаешься”. Думаю: “Ну так не может быть, так не бывает”» разметчики выбрали эмоции злости, грусти и удивления.

Причиной частой оценки текстовых фрагментов как эмоционально неоднородных может быть присущая тексту неоднозначность: не зная, с какой интонацией и мимикой была произнесена фраза, разметчик усматривал и оценивал в ней все возможные оттенки эмоций. Только 15 фрагментов (объектов разметки) были оценены как выражающие одни и те же смешанные эмоции по всем модальностям. Больше всего совпадений находим при оценке мультимодальных и текстовых фрагментов: 46,7% мультимодальных фрагментов со смешанными эмоциями были восприняты как таковые и при опоре только на вербальный канал, при этом практически во всех случаях набор эмоций, распознанных в двух условиях, был одинаков. Например, во фрагменте, где информантка говорит: «И для меня просто было невероятно осознать все то, что... это уже не тот образ моей сестры-одиннадцатиклассницы, она уже взрослая, ну, девушка, которая достаточно долгое время в браке, и вот-вот я стану тетей» и в полном фрагменте (аудио + видео), и только по тексту разметчики распознали две эмоции: радость и удивление. По-видимому, по критерию доступности маркеров эмоций для рецепции текстовая модальность ближе всего к условно «истинной оценке».

Наличие в разметке датасета не только эмоциональных категорий, но и оценок интенсивности позволяет получить информацию о том, как формировалось то или иное смешанное эмоциональное состояние. В качестве примера из нашего датасета рассмотрим высказывание «Мне не верится, я не выигрывала ни разу в жизни никаких билетов никуда, а тут вдруг выиграла». При разметке текстового фрагмента разметчики оценили «удивление» на 5 баллов и «радость» – на 4, поскольку «Мне не верится» однозначно указывает на удивление, а «выиграла» имеет положительную коннотацию. Видеофрагмент без звука получил только метку «радость» (3 балла), аудиофрагмент – по 4 балла для радости и удивления. Итоговыми метками мультимодального фрагмента также оказались удивление с баллом 3 и радость с баллом 2. На данном примере можно проследить, как формируется оценка доминирующей и дополнительной эмоции в зависимости от модальности: удивление при прочтении текста было оценено выше, чем при просмотре мультимодального фрагмента (5 баллов по тексту и 3 в полном фрагменте), а при просмотре видеоряда в мимике гово-

рящего разметчики вообще его не «считали», но почувствовали радость средней степени интенсивности (3 балла). Таким образом, мы можем наблюдать некоторые признаки того, что разные эмоции имеют различную степень маркированности в разных модальностях: одни эмоции тяготеют к выражению через вербальные маркеры, другие – через мимические или интонационные.

Эмоция vs модальность

Рассмотрим, как коррелировали между собой метки ведущих эмоций при разметке одних и тех же фрагментов, но представленных в разной модальности (рис. 7). Например, у нас есть объект разметки *X*, который был предъявлен трем разметчикам (разметчики и объекты в выборке для оценивания были разделены на 2 группы: 3 разметчика оценивали первую половину выборки, а еще 3 – вторую) в 4 модальностях и получил в каждой из них по 8 эмоциональных меток от каждого разметчика. Нас интересует процент случаев, когда метка

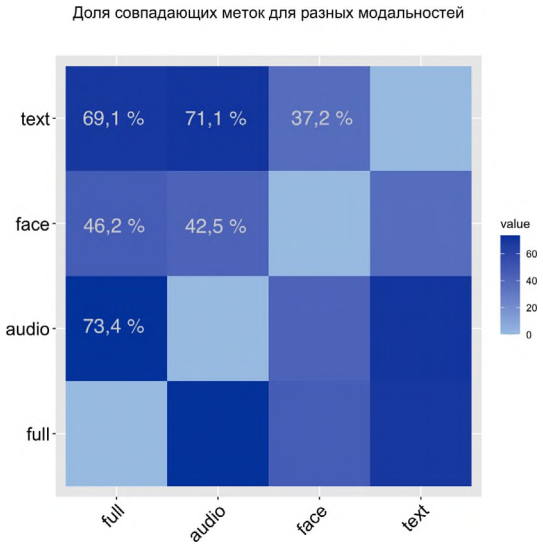

Рис. 7. Доля совпадающих меток для разных модальностей (подсчет по метке ведущей эмоции)

с самым большим весом по шкале эмоции совпала у всех разметчиков в каждой из модальностей предъявления объекта оценки.

Больше всего совпадений (73,4%) выявлено между оценками мультимодальных (полных) и звуковых фрагментов, а также звуковых и текстовых (71,1%). Считая, что метка мультимодального фрагмента наиболее полно отражает эмоции в нем, можем сделать вывод о том, что высокую значимость для распознавания эмоции имеет вербальный канал (т. е. читаемый текст), а при добавлении к нему просодического канала точность повышается еще на 4,3% (73,4% для аудио по сравнению 69,1% для письменного текста).

Отметив подобную общую закономерность, обратимся к частным случаям взаимодействия модальностей и их роли в узнавании разных эмоций разметчиками.

В табл. 4 представлены данные о том, каков процент (от общего числа объектов оценки, для которых данная эмоция была оценена разметчиками как ведущая) фрагментов, отмеченных данной эмоцией, при предъявлении в каждой из четырех модальностей. Так, для узнавания радости наиболее важной оказалась акустическая модальность: при оценке только по тексту данная эмоция выявлена разметчиками в

Таблица 4

Доля фрагментов, распознанных по каждой модальности, в % от общего числа фрагментов конкретной эмоции

	Аудио	Видеоряд	Текст
Радость	73,2	60,9	63,1
Грусть	62,7	49,2	62,7
Удивление	67,1	42,1	60,5
Злость	67,4	20,9	72,9
Страх	87,9	3,5	86,2
Отвращение	59,8	32,2	57,5
Нейтральная	78,4	59,5	71,4

63,1% фрагментов от общего числа «радостных», в то время как по аудио, когда к содержанию сообщаемого вербально добавилась интонация, – в 73,2% (см. табл. 4).

Рассмотрим пример из нашего датасета. Высказывание «Мы вспоминали, кто где живет. Заходили в магазины, в которые мы заходили после школы. И были на детской площадке, где мы катались на качелях» получило следующие оценки (эмоция – балл интенсивности):

– мультимодальный фрагмент: радость – 2;

– аудио: радость – 1;

– видеофрагмент: нейтральная – 5;

– текст: нейтральная – 5.

Так, в данном высказывании нет словесных маркеров, которые бы указали на переживание радости – будучи представлено в письменном виде, оно имеет нейтрально-повествовательный характер. Однако интонационная составляющая, а затем в дополнение к ней – мимическая в полном фрагменте позволили разметчикам идентифицировать эмоцию радости.

Аналогичную ситуацию наблюдаем для удивления: эмоция лучше всего считывается при представлении объекта оценки в аудиальной модальности: 67% на фоне 60% в текстовой модальности и 42,1% – в видео.

Злость же лучше идентифицируется по вербальным маркерам в текстовом представлении фрагмента 72,9% на фоне 67,2% – для аудио и 20,9 – для видео.

Для грусти и текстовая, и аудиальная модальность оказались равноценными – по 62,7% объектов оценки из категории «грусть» оказались распознаны в каждой из этих модальностей. Практически аналогичная ситуация для страха и отвращения при незначительном преобладании случаев их распознавания в аудиальной модальности над текстовой.

Однако при этом стоит отметить, что страх лучше всех других эмоций распознается по тексту и аудио – 86,2 и 87,9% соответственно. Это указывает на высокую значимость вербальных маркеров данной эмоции в нарративах: как при прослушивании аудио, так и при чтении текста высказывания разметчику был доступен вербальный канал (текст в устной или письменной форме), но если в случае с гру-

стью его доступность позволила узнать только 62% грустных фрагментов (а значит, при оценке мультимодального фрагмента значительный вклад внесли какие-то другие предикторы), то для страха вербальных маркеров было достаточно для узнавания более 80% объектов оценивания, отмеченных этой эмоцией.

Если рассмотреть отдельно колонку «Видеоряд» в табл. 4, то становится заметно, что по данной модальности зафиксирован наименьший процент идентификации всех эмоций. Тем не менее можно выделить лидеров и аутсайдеров: например, только по мимическим движениям лучше всего распознается радость (60,9%), а хуже всего – злость (20,9%) и страх (3,5%). Одно из возможных объяснений данного феномена состоит в том, что, когда участники рассказывали о чем-то радостном, они действительно погружались в это состояние и как будто заново его переживали, что ярко отражалось и в мимике, и в голосе, а не только вербально. В свою очередь, злость или страх – это негативные эмоции с высоким уровнем возбуждения. Вероятно, когда участники рассказывали страшные и раздражающие истории, эти эмоции заново переживались ими не с такой же силой и уровнем возбуждения, как в самой ситуации, о которой они рассказывали. В связи с чем эмоция сильнее и чаще отражалась вербально и меньше в мимике и просодии. Например, фрагмент «В общем, он никак не может доделать этот ремонт дурацкий» получил в среднем 3 балла по шкале злости как при оценке мультимодального фрагмента, так и текста, но невербальное поведение при произнесении данного высказывания оценивалось как нейтральное. По аудио была также определена злость, но она получила оценку в 2 балла: просодические характеристики как будто немного смягчили интенсивность словесного выражения эмоции, но оценка полного мультимодального фрагмента, когда разметчик и видел, и слышал говорящего, в итоге совпала именно с оценкой текста.

Стоит также отметить, что эмоция отвращения в целом мало проявилась во всех модальностях: 59,8% – аудио, 32,2% – видео, 57,5% – текст. Рассмотрим следующее высказывание: «И, ну, меня просто достало, что у меня постоянно холодильнике чем-то пахнет». В зависимости от оцениваемой модальности оно получило следующие оценки (эмоция – балл интенсивности):

– мультимодальный фрагмент: злость – 3, отвращение – 2;
– аудио: злость – 3;
– видеофрагмент: отвращение – 2;
– текст: злость – 3, отвращение – 2.

Так, в качестве ведущей эмоции фрагмента при оценке его в мультимодальной форме разметчики отметили злость, а в качестве дополнительной эмоции – отвращение. Аналогичные оценки получил и текст фрагмента. Это позволяет нам предположить, что ключевую роль сыграли вербальные маркеры, а именно слова «достало» и «пахнет». По оценкам аудио и видеоряда видно, как разные каналы формируют эмоцию: в дополнение к вербальному маркеру «пахнет» эмоция отвращения была поддержана мимикой, в то время как просодическое оформление больше указывало на злость. По-видимому, уже отмечавшаяся нами выше близость эмоций злости и отвращения приводит к слабой распознаваемости отвращения по отдельным модальностям. Только все модальности в комплексе помогают разметчикам идентифицировать данную эмоцию в полном мультимодальном фрагменте.

Итак, резюмируем результаты исследовательского анализа полученного датасета.

1. Предложенная процедура индуцирования эмоциональных состояний наилучшим образом подошла для провоцирования эмоций радости, грусти и злости, но не позволила получить большое количество материала для эмоций страха и удивления.

2. Наибольшая согласованность аннотаторов наблюдается при разметке полных (мультимодальных) объектов оценивания или при их предъявлении в текстовой модальности; наименьшая – при предъявлении в формате видео (без звука).

3. Наибольшая согласованность по отдельным эмоциональным классам была достигнута при предъявлении мультимодальных фрагментов для эмоций радости и грусти, а наименьшая – в случаях смешанных негативных эмоций, когда разметчики колебались между грустью, злостью и отвращением.

4. Анализ объектов оценки, получивших в итоге несколько эмоциональных меток, показал, что разные модальности вносят разный вклад в идентификацию той или иной эмоции. Эта же тенденция про-

явилась и при анализе доли правильно идентифицированных (относительно разметки полных (мультимодальных) объектов оценивания) эмоций при предъявлении тех же объектов оценивания в отдельных модальностях: текстовой, аудио и видео (без звука). Таким образом, выявлены тенденции к преимущественной идентификации эмоций радости и удивления при предъявлении объекта оценивания в аудиоформате; грусти, страха и отвращения – в аудио и текстовой модальностях; злости – в текстовой модальности. Однако при этом страх точнее всех остальных эмоций идентифицируется по аудио и тексту, а отвращение – хуже всего поддается идентификации независимо от модальности. Предъявление объекта оценивания в видеоформате снижает процент точного распознавания для всех эмоций, но в наименьшей степени – для радости, а в наибольшей – для страха.

Следовательно, мы можем заключить, что точность автоматической эмоциональной классификации речевого материала на русском языке для большинства эмоций должна быть выше при использовании моделей, обученных для анализа текстов (транскриптов). При использовании же моделей, обученных для распознавания эмоций по мимическим движениям, скорее всего, низкую точность распознавания будут иметь страх и злость.

Мы решили использовать полученные нами агрегированные оценки для полных (мультимодальных) объектов оценивания, собранных в нашем датасете, в качестве «золотого стандарта» для оценивания качества эмоциональной классификации существующих моделей.

Кейс практического применения датасета для оценки качества существующих моделей эмоциональной атрибуции текстов

Было выбрано 8 моделей:

– для анализа эмоций по мимике:

1) hsemotion, https://github.com/av-savchenko/face-emotion-recognition?tab=readme-ov-file;

– для анализа эмоций в тексте:

2) дообученная на задачу распознавания эмоций модель RuBERT tiny (https://huggingface.co/cointegrated/rubert-tiny2);

3) дообученная на задачу распознавания эмоций модель RuBERT Base (https://huggingface.co/seara/rubert-base-cased-cedr-russian-emotion);

4) модель из библиотеки aniemore (также на основе архитектуры RuBERT tiny, https://huggingface.co/Aniemore/rubert-tiny2-russian-emotion-detection);

– для анализа эмоций в аудио:

5) дообученная на задачу распознавания эмоций модель HuBERT (https://huggingface.co/xbgoose/hubert-large-speech-emotion-recognition-russian-dusha-finetuned);

6) дообученная на задачу распознавания эмоций модель wav2vec-XLS-R (https://huggingface.co/KELONMYOSA/wav2vec2-xls-r-300m-emotion-ru);

7) модель из пакета aniemore (архитектура wavlm, https://huggingface.co/Aniemore/wavlm-emotion-russian-resd);

– и одна мультимодальная модель (текст + аудиосигнал):

8) модель из пакета aniemore (https://huggingface.co/Aniemore/wavlm-bert-fusion-s-emotion-russian-resd).

Мы получили предсказания моделей для записей нашего датасета, после чего сравнили их с аннотациями разметчиков. Для удобства сравнения использовались не вероятности, а словесные метки эмоций (табл. 5).

Таблица 5

Доля совпадений между предсказаниями моделей и аннотациями разметчиков

Модель	% совпадений с аннотациями разметчиков
hsemotion_video	25,63
RuBERT_tiny_text	50,89
RuBERT_base_text	54,11
aniemore_text	58,39
HuBERT_voice	39,59
wav2vec-XLS-R_voice	39,02
aniemore_voice	17,92
aniemore_multimodal	16,96

Лучше всего себя показали текстовые модели: в 50–58% случаев их предсказания совпадают с оценками разметчиков. Наилучший результат (58,39%) показала модель, которая дообучалась на схожем с нашим материале (реплики диалогов, голосовые сообщения). Две акустические модели (HuBERT_voice и wav2vec-XLS-R_voice) показали примерно одинаковый результат: около 40% совпадений с человеческими оценками. Сходство результатов можно объяснить тем, что обе модели дообучались на материале датасета DUSHA [30]. Третья же голосовая модель (aniemore_voice) дообучалась на ином материале (голосовые сообщения, записи диалогов и отдельных фраз): в отличие от датасета DUSHA, где часть записей сделана с привлечением актеров (сыгранные эмоции), а другая часть – это естественный материал (подкасты), в обучающей выборке модели aniemore весь материал «сыгранный». Это еще раз указывает на необходимость использования аутентичного материала в обучении моделей распознавания эмоций.

Модель распознавания эмоций по мимике совпала в своих предсказаниях с оценками разметчиков в 25,63% случаев, однако такой результат прогнозируем, поскольку процесс разметки показал, что при опоре только на визуальный канал даже между разметчиками-людьми не достигается высокая степень согласия. Иными словами, только визуального канала недостаточно для точного определения эмоции.

Заключение

Созданный датасет представляет собой валидный источник данных о распознавании эмоций носителями русского языка в речевых фрагментах, поданных в трех отдельных модальностях (аудио, текст, видео), а также в мультимодальном (полном) формате, метки которого рассматривались как априори истинные. Разработанный нами метод индуцирования эмоционального состояния доказал свою состоятельность и в дальнейшем может быть использован другими исследователями. Проведенная с использованием датасета исследовательская работа подтвердила несколько уже устоявшихся в психологии эмоций и лингвоэмотиологии мнений, например, суждение о том, что грусть и радость представляют собой некоторую базовую или прототипичес-

кую оппозицию эмоций – исследование показало, что согласованность мнений разметчиков для этих двух эмоций выше всего. Кроме того, сравнение точности распознавания эмоций аннотаторами (по сравнению с результатами разметки мультимодальных фрагментов) при предъявлении речевых фрагментов в разных модальностях также согласуется с наблюдениями о том, что текстовая модальность содержит максимум информации для идентификации большинства эмоций. В то же время мы получили новые данные о неравнозначной роли модальностей для идентификации разных эмоций, а также о том, что негативные эмоции чаще проявляются одновременно и совместно, формируя смешанную эмоцию. Практическое применение результатов разметки, систематизированных в датасете, продемонстрировало, что датасет может быть с успехом использован как «золотой стандарт» для оценки качества автоматического эмоционального анализа речи на русском языке: модели, оценивающие тексты, имеют большую точность распознавания, нежели, например, оценивающие эмоцию по аудиоданным или мимическим движениям в видеозаписях.

Вместе с тем мы видим и ограничения работы. Прежде всего, они связаны с точностью границ выделения того или иного объекта оценивания (речевого фрагмента). Достаточно трудно экспертно определить, где заканчивается одна эмоция и начинается другая. Четких идентифицируемых автоматических критериев нам не удалось подобрать.

В дальнейшем необходимо продолжить работу по оценке качества существующих моделей, сфокусировавшись на отдельных классах эмоций, отдельно поработать со смешанными эмоциями. Кроме того, привлекательным представляется дообучение существующих моделей или их тонкая настройка с использованием нашего датасета. В последнее время большую популярность получило использование Больших языковых моделей для решения в том числе задач эмоционального анализа через различные стратегии промптинга, т. е. специально сформулированные инструкции, адресуемые БЯМ. Мы планируем провести и такое тестирование возможностей датасета.

На примере описания структуры и возможностей датасета мы предприняли попытку обосновать тезис о том, что в новых технологичных условиях работы лингвиста именно датасет становится одной

из возможных форм реализации экспертного знания специалиста, тем продуктом, который, наряду с традиционным словарем, ждет от лингвиста сообщество. «Наивный» носитель языка, стремясь делегировать свою интерпретирующую функцию предобученным моделям или БЯМ, нуждается не только в парах «лемма – толкование», но и в диаде «текстовые примеры – категориальная метка, помогающая интерпретировать». Иными словами, датасет может сегодня рассматриваться как новая ипостась словаря.

Список источников

1. *Шаховский В.И.* Обоснование лингвистической теории эмоций // Вопросы психолингвистики. 2019. № 1. С. 22–37.
2. *Бабенко Л.Г.* Лингвопсихология на методологической базе когнитивистики: лексикографический аспект // Изв. Урал. федер. ун-та. Сер. 2: Гуманитар. науки. 2020. Т. 22, № 3 (200). С. 264–278.
3. *Бабенко Л.Г.* Алфавит эмоций: словарь-тезаурус эмотивной лексики. Екатеринбург; Москва : Кабинетный ученый, 2020. 431 с.
4. *История* русской лексикографии / отв. ред. Ф.П. Сороколетов. СПб. : Наука, 1998. 610 с.
5. *Кибрик А.А.* Мультимодальная лингвистика // Когнитивные исследования: сборник научных трудов. Вып. 4. М. : Издательство Института психологии РАН, 2010. С. 135–152.
6. *Ирисханова О.К.* Полимодальный дискурс как объект исследования // Полимодальные измерения дискурса / отв. ред. О.К. Ирисханова. М. : ЯСК, 2021. С. 15–33.
7. *The Routledge* Handbook of Multimodal Analysis / ed. by C. Jewitt. Routledge handbooks, 2016. 527 p.
8. *Pan B., Hirota K., Jia Z., Dai Y.* A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods // Neurocomputing. 2023. V. 561. P. 126866. doi: 10.1016/j.neucom.2023.126866
9. *Mlakar I., Kačič Z., Rojc M.* A Corpus for investigating the multimodal nature of multi-speaker spontaneous conversations – EVA Corpus // WSEAS Transactions on Information Science and Applications. 2017. V. 14. P. 213–226.
10. *Das A., Sarma M.S., Hoque M.M., Siddique N., Dewan M.A.A.* AVaTER: Fusing Audio, Visual, and Textual Modalities Using Cross-Modal Attention for Emotion Recognition // Sensors (Basel). 2024. V. 24 (18). P. 5862. doi: 10.3390/s24185862. PMID: 39338607; PMCID: PMC11436096

11. *Perepelkina O., Kazimirova E., Konstantinova M.* RAMAS: Russian Multimodal Corpus of Dyadic Interaction for studying emotion recognition: PeerJ Preprints. 2018. V. 6. doi: 10.7287/PEERJ.PREPRINTS.26688V1
12. *Poria S., Hazarika D., Majumder N., Naik G., Cambria E., Mihalcea R.* Meld: A multimodal multi-party dataset for emotion recognition in conversations // arXiv. 2018. doi: 10.48550/arXiv.1810.02508
13. *Zadeh A.B., Liang P.P., Poria S.E., Morency L.P.* Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018. V. 1. P. 2236–2246.
14. *Zhao J., Zhang T., Hu J., Liu Y., Jin Q., Wang X., Li H.* Multi-modal multi-scene multi-label emotional dialogue database // arXiv. 2022. doi: 10.48550/arXiv.2205.10237
15. *Zhalehpour S., Onder O., Akhtar Z., Erdem C.E.* BAUM-1: A Spontaneous Audio-Visual Face Database of Affective and Mental States // IEEE Transactions on Affective Computing. 2017. V. 8 (3). P. 300–313.
16. *Parada-Cabaleiro E., Costantini G., Batliner A., Schmitt M., Schuller B.W.* DEMoS – An Italian Emotional Speech Corpus – Elicitation methods, machine learning, and perception // Language Resources and Evaluation. 2020. V. 54. P. 341–383.
17. *Kotov A., Budyanskaya E.* The Russian emotional corpus: communication in natural emotional situations // Компьютерная лингвистика и интеллектуальные технологии: материалы междунар. конф. «Диалог». Вып. 11(18): в 2 т. Т. 1: Основная программа конференции. М. : Изд-во РГГУ, 2012. С. 296–307.
18. *Ekman P., Davidson R.J.* The nature of emotion: Fundamental questions. Oxford : Oxford University Press, 1994. 496 p.
19. *Tomkins S.S.* PAT Interpretation: Scope and Technique. New York : Springer Publishing, 1959. 18 p.
20. *Plutchik R.* The Nature of Emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice // American Scientist. 2001. V. 89, No. 4. P. 344–350.
21. *Изард К.Э.* Психология эмоций. СПб. : Питер, 2006. 464 с.
22. *Wang F., Yu J., and Xia R.* Generative Emotion Cause Triplet Extraction in Conversations with Commonsense Knowledge // Findings of the Association for Computational Linguistics: EMNLP 2023. 2023. P. 3952–3963.
23. *Ringeval F., Sonderegger A., Sauer J., Lalanne D.* Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions // 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2013. 2013. P. 1–8.

24. *Busso C., Bulut M., Lee C., Kazemzadeh A., Mower E., Kim S., Chang J.N., Lee S., Narayanan S.* IEMOCAP: interactive emotional dyadic motion capture database // Language Resources and Evaluation. 2008. V. 42 (4). P. 335–359.
25. *Tzirakis P., Zafeiriou S., Schuller B.* Real-world automatic continuous affect recognition from audiovisual signals // Computer Vision and Pattern Recognition, Multimodal Behavior Analysis in the Wild / Ed. by X. Alameda-Pineda, E. Ricci, N. Sebe. Academic Press, 2019. P. 387–406.
26. *Hayes A.F., Krippendorff K.* Answering the call for a standard reliability measure for coding data // Communication Methods and Measures. 2007. V. 1. P. 77−89.
27. *Mills C., D'Mello S.* On the validity of the autobiographical emotional memory task for emotion induction // PLoS One. 2014. V. 9 (4). pmid:24776697
28. *Tkachenko M., Malyuk M., Holmanyuk A., Liubimov N.* Label Studio: Data labeling software. 2020. Available from: https://github.com/heartexlabs/label-studio
29. *Старостина Е.Г., Тэйлор Г.Д., Квилти Л.К., Бобров А.Е., Мошняга Е.Н., Пузырева Н.В., Боброва М.А., Ивашкина М.Г., Кривчикова М.Н., Шаврикова Е.П., Бэгби М.* Торонтская шкала алекситимии (20 пунктов): валидизация русскоязычной версии на выборке терапевтических больных // Социальная и клиническая психиатрия. 2010. № 20(4). С. 31–38.
30. *KondratenkoV., Sokolov A., Karpov N., Kutuzov O., Savushkin N., Minkin F.* Large Raw Emotional Dataset with Aggregation Mechanism (Version 1) // arXiv. 2022. doi: 10.48550/ARXIV.2212.12266

References

1. Shakhovskiy, V.I. (2019) Obosnovanie lingvisticheskoy teorii emotsiy [Justification of the linguistic theory of emotions]. *Voprosy psikholingvistiki*. 1. pp. 22–37.
2. Babenko, L.G. (2020) Linguopsychology Based on the Methods of Cognitive Studies: The Lexicographic Aspect. *Izvestiya Ural'skogo federal'nogo universiteta. Ser. 2: Gumanitarnye Nauki*. 22 (3). pp. 264–278. (In Russian). doi: 10.15826/izv2.2020.22.3.057
3. Babenko, L.G. (2020) *Alfavit emotsiy: slovar'-tezaurus emotivnoy leksiki* [Alphabet of emotions: a dictionary-thesaurus of emotive vocabulary]. Yekaterinburg–Moscow: Kabinetnyy uchenyy.
4. Sorokoletov, F.P. (ed.) (1998) *Istoriya russkoy leksikografii* [History of Russian Lexicography]. St. Petersburg: Nauka.
5. Kibrik, A.A. (2010) Mul'timodal'naya lingvistika [Multimodal linguistics]. In: *Kognitivnye issledovaniya: sbornik nauchnykh trudov*. 4. Moscow: RAS Institute of Psychology. pp. 135–152.
6. Iriskhanova, O.K. (2021) Polimodal'nyy diskurs kak ob"ekt issledovaniya [Polimodal discourse as an object of research]. In: *Polimodal'nye izmereniya diskursa* [Polymodal dimensions of discourse]. Moscow: YASK. pp. 15–33.

7. Jewitt, C. (ed.) (2016) *The Routledge Handbook of Multimodal Analysis.* Routledge Handbooks.
8. Pan, B., Hirota, K., Jia, Z. & Dai, Y. (2023) A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods. *Neurocomputing*. 561. p. 126866. doi: 10.1016/j.neucom.2023.126866
9. Mlakar, I., Kačič, Z. & Rojc, M. (2017) A Corpus for investigating the multimodal nature of multi-speaker spontaneous conversations – EVA Corpus. *WSEAS Transactions on Information Science and Applications*. 14. pp. 213–226.
10. Das, A., Sarma, M.S., Hoque, M.M., Siddique, N. & Dewan, M.A.A. (2024) AVaTER: Fusing Audio, Visual, and Textual Modalities Using Cross-Modal Attention for Emotion Recognition. *Sensors* (Basel). 24 (18). p. 5862. doi: 10.3390/s24185862
11. Perepelkina, O., Kazimirova, E. & Konstantinova, M. (2018) RAMAS: Russian Multimodal Corpus of Dyadic Interaction for studying emotion recognition. *PeerJ Preprints*. 6. doi: 10.7287/peerj.preprints.26688v1
12. Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., Mihalcea, R. (2018) Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv*. doi: 10.48550/arXiv.1810.02508
13. Zadeh, A.B., Liang, P.P., Poria, S. & Morency, L.P. (2018) Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. 1. pp. 2236–2246.
14. Zhao, J., Zhang, T., Hu, J., Liu, Y., Jin, Q., Wang, X. & Li, H. (2022) Multi-modal multi-scene multi-label emotional dialogue database. *arXiv*. doi: 10.48550/arXiv.2205.10237
15. Zhalehpour, S., Onder, O., Akhtar, Z. & Erdem, C.E. (2017) BAUM-1: A Spontaneous Audio-Visual Face Database of Affective and Mental States. *IEEE Transactions on Affective Computing*. 8 (3). pp. 300–313.
16. Parada-Cabaleiro, E., Costantini, G., Batliner, A., Schmitt, M. & Schuller, B.W. (2020) DEMoS – An Italian Emotional Speech Corpus – Elicitation methods, machine learning, and perception. *Language Resources and Evaluation*. 54. pp. 341–383.
17. Kotov, A. & Budyanskaya, E. (2012) The Russian emotional corpus: communication in natural emotional situations. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii: Dialog*. 11 (18). Moscow: RSUH. pp. 296–307.
18. Ekman, P. & Davidson, R.J. (1994) *The Nature of Emotion: Fundamental Questions*. Oxford: Oxford University Press.
19. Tomkins, S.S. (1959) *PAT Interpretation: Scope and Technique*. New York: Springer Publishing.

20. Plutchik, R. (2001) The Nature of Emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*. 89 (4). pp. 344–350.
21. Izard, K.E. (2006) *Psikhologiya emotsiy* [Psychology of emotions]. St. Petersburg: Piter.
22. Wang, F., Yu, J. & Xia, R. (2023) Generative Emotion Cause Triplet Extraction in Conversations with Commonsense Knowledge. *Findings of the Association for Computational Linguistics: EMNLP.* 2023. pp. 3952–3963.
23. Ringeval, F., Sonderegger, A., Sauer, J. & Lalanne, D. (2013) Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2013*. pp. 1–8.
24. Busso, C., Bulut, M., Lee, C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S. & Narayanan, S. (2008) IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*. 42 (4). pp. 335–359.
25. Tzirakis, P., Zafeiriou, S. & Schuller, B. (2019) Real-world automatic continuous affect recognition from audiovisual signals. In: Alameda-Pineda, X., Ricci, E., Sebe, N. (eds) *Computer Vision and Pattern Recognition, Multimodal Behavior Analysis in the Wild*. Academic Press. pp. 387–406.
26. Hayes, A.F. & Krippendorff, K. (2007) Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*. 1. pp. 77–89.
27. Mills, C. & D'Mello, S. (2014) On the validity of the autobiographical emotional memory task for emotion induction. *PLoS One*. 9 (4). doi: 10.1371/journal.pone.0095837
28. Tkachenko, M., Malyuk, M., Holmanyuk, A. & Liubimov, N. (2020) *Label Studio*: *Data labeling software*. [Online] Available from: https://github.com/heartexlabs/label-studio
29. Starostina, E.G., Taylor, G.D., Quilty, L.C., Bobrov, A.E., Moshnyaga, E.N., Puzyreva, N.V., Bobrova, M.A., Ivashkina, M.G., Krivchikova, M.N., Shavrikova, E.P. & Bagby, M. (2010) Torontskaya shkala aleksitimii (20 punktov): validizatsiya russkoyazychnoy versii na vyborke terapevticheskikh bol'nykh [Toronto Alexithymia Scale (20 items): validation of the Russian version on a sample of therapeutic patients]. *Sotsial'naya i klinicheskaya psikhiatriya*. 20 (4). pp. 31–38.
30. Kondratenko, V., Sokolov, A., Karpov, N., Kutuzov, O., Savushkin, N. & Minkin, F. (2022) Large Raw Emotional Dataset with Aggregation Mechanism (Version 1). *arXiv*. doi: 10.48550/arXiv.2212.12266

Сведения об авторах:
Колмогорова Анастасия Владимировна – д-р филол. наук, проф., зав. лабораторией языковой конвергенции НИУ ВШЭ в Санкт-Петербурге (Санкт-Петербург, Россия). E-mail: akolmogorova@hse.ru

Куликова Елизавета Романовна – младший научный сотрудник лаборатории языковой конвергенции НИУ ВШЭ в Санкт-Петербурге (Санкт-Петербург, Россия). E-mail: Kulikova.E.R@hse.ru

Авторы заявляют об отсутствии конфликта интересов.

Information about the authors:
Anastasia V. Kolmogorova, Dr. Sci. (Philology), full professor, head of the Linguistic Convergence Laboratory, National Research University Higher School of Economics (Saint Petersburg, Russian Federation). E-mail: akolmogorova@hse.ru
Elizaveta R. Kulikova, junior researcher, National Research University Higher School of Economics (Saint Petersburg, Russian Federation). E-mail: Kulikova.E.R@hse.ru

The authors declare no conflicts of interests.

Статья поступила в редакцию 26.02.2025;
одобрена после рецензирования 28.04.2025; принята к публикации 23.05.2025

The article was submitted 26.02.2025;
approved after reviewing 28.04.2025; accepted for publication 23.05.2025