

Приложение к итоговому отчету по проекту РФФИ № 18-29-22061 «Разработка комплексного подхода к анализу русскоязычных аудио-, видео- и текстовых материалов с целью выявления агрессивного поведения пользователей информационного пространства на основе нейросетевых технологий»

Целью проекта является разработка комплексного подхода к семантическому анализу мультимедийных материалов пользователя (аудио-, видео-, текстовые сообщения) в Интернет-пространстве и определению эмоциональной окрашенности сообщений, передаваемых посредством различных модальностей, позволяющего выявить потенциально деструктивное информационное воздействие и негативные личностные тенденции. В ходе разработки такого подхода предполагалось использование опыта не только современных исследований в области нейронных сетей, обработки сигналов, но и в области эмотиологии (лингвистика эмоций), что в совокупности позволяет выявлять агрессивное поведение пользователей информационного пространства на основе одной или нескольких модальностей. Разработка данного подхода включает в себя следующие задачи:

- разработка методики выявления агрессивного поведения пользователя в информационном пространстве на основе многомодального анализа мультимедийных материалов пользователя;
- разработка алгоритма кластеризации разнородных материалов на семантические подгруппы с целью комплексного подхода к анализу мультимедийных материалов пользователя;
- разработка архитектуры многозадачной нейронной сети для многомодального анализа разнородных материалов пользователя с целью выявления признаков, характерных для агрессивного поведения пользователя;
- разработка и тестирование прототипа системы выявления агрессивного поведения пользователя в информационном пространстве.

На первом этапе выполнения проекта был проведен обзор современного состояния исследований в области психологии: рассмотрены существующие определения агрессии, множественность понятия агрессии, проведено разграничение понятий агрессии, гнева, враждебности и насилия, деструктивного поведения. Рассмотрение существующих классификаций видов агрессии показало, что с точки зрения анализа поведения пользователя в информационном пространстве необходимо рассматривать, в первую очередь, вербальные, пассивные и косвенные проявления агрессии. Рассмотрены вопросы регистрации, идентификации и анализа агрессивного поведения с точки зрения

возможности автоматизации данных процессов. Многообразие средств информационного обмена в интернет-пространстве обуславливает необходимость работы как с текстовыми материалами и изображениями, так и с аудио- и видеозаписями. Для этого проанализированы существующие методы и признаки выявления потенциально деструктивного поведения пользователей информационного пространства на основе различных модальностей. По результатам обзора выделен набор признаков, характерных для агрессивного поведения пользователя в информационном пространстве с учетом различных модальностей. Так, с точки зрения анализа текстовых материалов рассмотрены лексические и дискурсивные языковые средства, используемые при агрессивном речевом поведении. С точки зрения анализа видеоматериалов и изображений целесообразно оценивать: мимику, жесты, движения тела, расстояние между коммуникантами. С точки зрения анализа аудиоданных целесообразно рассматривать: модуляционные характеристики речи (речевые мелодические контуры, ритм, темп, паузы) и словесные средства (семантика, лексика, грамматика, синтаксис, стилистика). Кроме того, при анализе изображений, аудио- и видеоматериалов необходимо определить наличие встроенных неблагоприятных для человека шумов, неприятных для глаза изображений с целью исключения влияния на физическое состояние человека.

В ходе второго этапа выполнения проекта было проанализировано текущее состояние исследований в области определения психофизиологического состояния пользователя информационного пространства на основе текстовых, аудио- и видеоматериалов. Для определения психофизиологического состояния пользователя по текстовым материалам используются методы анализа эмоциональной окраски сообщений (sentiment analysis), а также типовые схемы обработки текста, включающие удаление стоп-слов, токенизацию, разметку частей речи и т. д., и типовые способы извлечения признаков, которые формируют векторное представление текста (например, «bag of words» или «bag of stems»). Проанализированы особенности представления разнородных материалов в информационном пространстве. В частности, для интернет-пространства характерно: использование смеси текстового и аудиовизуального контента, смеси разных языков, специфической лексики, наличие сообщений с ошибками, опечатками, нетипичной пунктуацией. Могут использоваться приемы для обхода автомодерации. Все это затрудняет применение классических методов обработки естественного языка.

При анализе психофизиологического состояния пользователя по аудиоматериалам были выявлены следующие основные недостатки существующих методов: зачастую авторы разрабатывают свою специфичную модель эмоций, применительно к которой определяют

информативные признаки для классификации эмоций; не существует общепринятой модели эмоций в области звучащей речи; не существует универсального информативного и емкого набора признаков, отражающего изменение эмоций диктора; большое количество интерпретаций звукового сигнала не позволяет добиться однозначности при его анализе; разрабатываемые методы являются узконаправленными и их редко можно интегрировать в другие задачи.

При анализе методов оценки психофизиологического состояния пользователя по изображениям выявлено, что ни один из существующих методов не обеспечивает высокую точность определения всех 7 базовых эмоций, представляемых в классических базах данных. Наибольшая проблема наблюдается в выявлении нейтрального, спокойного состояния, а также отделения счастья, удивления и злости.

В контексте исследования подходов к распознаванию агрессии и ее проявлений, основывающихся на анализе динамики кадров видеопотока, был сформирован набор признаков, характерных для деструктивного поведения пользователя в информационном пространстве в виде электронной таблицы численных значений ряда выделенных признаков, полученных с помощью анализа потоков изображений лица с изменением эмоции от нейтральной к деструктивной. Обзор показал, что в задачах распознавания агрессивного поведения существует ряд проблем: известные модели и методы, демонстрирующие высокие результаты распознавания, требуют параллельной обработки RGB-кадров изображения и кадров оптического потока или же сложной предварительной обработки RGB-кадров видеопотока, и соответственно имеют высокую вычислительную сложность; наблюдается недостаток репрезентативных размеченных баз видеоданных для обучения, на которых представлены агрессивные действия.

В ходе выполнения проекта была разработана архитектура нейронной сети для анализа видеоряда с целью выявления признаков, характерных для агрессивного поведения пользователей, по лицам пользователей, представленным на кадрах видео. Предложенная архитектура основывается комбинировании нейронных сетей, выполняющих анализ пространственных признаков визуальных объектов на кадрах видеопоследовательности, а также нейронных сетей, выполняющих временной анализ динамики этих признаков, и применении подхода переноса обучения. Архитектура отличается использованием каскада, состоящего из двух элементов – глубокой сверточной нейронной сети, предварительно обученной задаче распознавания мимических проявлений по изображениям лиц на большом репрезентативном наборе данных и выполняющей выделение наиболее значимых признаков представлений пространственных составляющих мимических проявлений на

отдельных кадрах видео, а также трехмерной сверточной нейронной сети, которая выполняет пространственно-временной анализ выделенных признаков. При этом, временная составляющая анализа позволяет выделить наиболее значимые динамические характеристики и повысить точность распознавания мимических изменений в динамике, в то время как пространственная составляющая обработки кадров трехмерными сверточными нейронными сетями анализирует и выделяет дополнительную пространственную информацию, что также положительно сказывается на итоговых результатах распознавания. В результате апробации данной модели на наборе русскоязычных данных RAMAS был получен результат 76.9% вернораспознанных экземпляров, который превзошел результат (74.4%), показанный трехмерной сверточной нейронной сетью архитектуры R3D, не обучавшейся предварительно задаче распознавания выражений лиц.

Разработана архитектура нейронной сети для анализа видеоряда с целью выявления признаков, характерных для агрессивного поведения пользователей, по двигательной активности тела пользователей, представленной на кадрах видео. Предложенная архитектура основывается на применении подхода переноса обучения, применении глубоких трехмерных сверточных нейронных сетей для анализа пространственно-временных признаков двигательной активности тела человека в видеопотоке, а также добавлении дополнительных промежуточных нейросетевых модулей, выполняющих выделение наиболее значимой информации. Архитектура отличается использованием промежуточных модулей, расположенных между сверточными блоками глубоких трехмерных сверточных нейросетей и основанных на механизме нейросетевого внимания, который, в свою очередь, выполняет анализ признаков представлений, генерируемых промежуточными слоями глубоких нейронных сетей, и выделяет наиболее значимые составляющие этих промежуточных представлений. Применение блоков нейросетевого внимания позволяет существенно повысить эффективность распознавания двигательной активности, которая обычно сопутствует агрессивному поведению. В результате апробации данной нейросетевой архитектуры на наборе русскоязычных данных RAMAS был достигнут результат, равный 80.4% вернораспознанных экземпляров. Данный результат превзошел аналогичный показатель у аналогичной трехмерной сверточной нейросетевой модели, но без механизма нейросетевого внимания, который составил 78.8% верно распознанных экземпляров.

Коллективом исполнителей проекта была разработана архитектура нейронной сети для анализа аудиосигнала речи пользователей информационного пространства, основанная

на применении глубоких сверточных нейронных сетей для обработки аудиосигнала, представленного в виде логарифмированной мел-спектрограммы. Предлагаемая архитектура отличается применением подхода многозадачного обучения, заключающегося в анализе и извлечении высокоуровневых признаков аудиосигнала и одновременным выполнением задач распознавания эмоций, личности спикера и пола спикера. Выбор данного подхода основывается на следующих его достоинствах: использование индуктивного переноса знаний между задачами в качестве элемента регуляризации, из чего следует возможность извлечения более высокоуровневой и надежной признаковой информации из представлений при меньших объемах тренировочных данных; а также сниженное количество параметров и ресурсных затрат на обучение по сравнению с несколькими обученными решением отдельных задач моделями. Предлагаемый подход отличается от известных применением автоматического взвешивания значений функции потерь отдельных задач на каждом пакете обрабатываемых данных в процессе обучения методом взвешенного среднего, что позволяет урегулировать скорость обучения каждой задачи и улучшить тем самым процесс индуктивного переноса знаний, что в свою очередь приводит к лучшей обобщающей способности модели. В результате апробации данной нейросетевой модели на русскоязычном наборе данных RAMAS был достигнут 91.5% верно выполненных распознаваний у нейронной сети архитектуры VGG. Данный результат существенно превзошел результат аналогичной модели VGG, у которой подход многозадачного обучения не применялся (74.1% верных распознаваний).

Разработана архитектура нейронной сети для анализа текста, генерируемого пользователем, с целью распознавания агрессивного поведения в информационном пространстве. Предлагаемая архитектура отличается применением специальной аугментации данных при обучении искусственной сети. Аугментация состоит в дополнении исходных данных маркерами, отражающими лексические, морфологические, статистические, косвенные признаки агрессии. Использование данного метода обусловлено необходимостью комплексного подхода к выявлению агрессивных русскоязычных текстов, который бы использовал возможности машинного обучения, учитывал лексические, синтаксические, грамматические особенности русского языка и позволял использовать широкий круг датасетов с многомодальным контентом. Отличием предложенного метода является использование построенной классификации признаков агрессии в тексте для аугментации исходных данных, что позволяет повысить точность выявления агрессивного русскоязычного текстового контента до 85%, 87%, 71% для наборов русскоязычных текстовых данных RTLC, RuTwwetCorp и RAMAS соответственно. Данные результаты превзошли аналогичные результаты той же самой архитектуры, в которой не используется

аугментация (80%, 83%, 68% для наборов русскоязычных текстовых данных RTLC, RuTwwetCorp и RAMAS соответственно).

В ходе выполнения проекта разработана архитектура нейронной сети для выделения и анализа комбинированных признаков агрессивного поведения, полученных с помощью разработанных ранее архитектур нейронных сетей, которые выполняют обработку отдельных модальностей. Предложенная архитектура отличается новым способом связывания векторов признаков отдельных модальностей для формирования единого признакового пространства, которое инвариантно виду обрабатываемой модальности и зависит только от распознаваемого класса. Этот способ заключается в обработке промежуточных векторов признаков отдельных модальностей одной и той же глубокой нейронной сетью. При этом глубина нейронной сети, с одной стороны, и возможность обработки сложных нелинейных зависимостей, с другой, обеспечивают необходимую инвариантность распознавания относительно класса. Другой отличительной особенностью предлагаемой архитектуры является новый алгоритм обучения нейронной сети, который основывается на пакетном градиентном спуске, формировании обучающих пакетов данных из векторов признаков всех обрабатываемых модальностей, одновременной обработкой нейронной сетью всего пакета векторов признаков, усреднении выходных ошибок для всего обработанного пакета данных. Именно это усреднение позволяет нейронной сети эффективно настраивать свои параметры таким образом, что наиболее значимой информацией становится информация об искомым признаках агрессивного поведения, содержащихся в векторах признаков отдельных модальностей, и обеспечивается инвариантность относительно вида обрабатываемой модальности. Применение таких подходов обеспечивает повышение эффективности распознавания агрессивного поведения в социкиберфизических системах. Кроме того, предлагаемая архитектура позволяет работать в условиях недостатка или отсутствия информации отдельных модальностей, что может повысить устойчивость функционирования систем распознавания агрессии, в основе которых лежит эта нейросетевая архитектура. В результате апробации данной нейросетевой модели на русскоязычном многомодальном наборе данных RAMAS был достигнут результат доли верно выполненных распознаваний, равный 94.3%, что превзошло аналогичный показатель у модели, выполняющей классификацию экземпляров отдельных модальностей и использующей простое усреднение результатов этой классификации (92.7%), и у модели, выполняющей ранее связывание отдельных модальностей (92.4%).

Полученные за третий период важнейшие научные результаты (с описанием их оригинальности и уникальности, а также значимости для развития междисциплинарных направлений, и их соответствие задачам темы исследований)

1. Разработана классификация агрессивного поведения, проявляемого на видео, пользователей информационного пространства в различных ситуациях. Классификация основывается на общепсихологическом определении агрессии, содержащем наиболее значимые признаки этого феномена (направленный и деструктивный характер действий), а также одной из наиболее распространённых классификаций агрессии, данной А. Бассом (основания классификации: модальность проявления – физическая/вербальная, активность субъекта – активная/пассивная, открытость действий – прямая/косвенная). Главными особенностями разработанной классификации являются: учет специфики агрессивного поведения пользователей информационного пространства, в котором распознавание некоторых видов агрессии (пассивная и косвенная) по классификации А. Басса существенно затруднено; явное выделение объектов нападения; разделение агрессивного поведения и сопутствующих эмоциональных и двигательных реакций; рассмотрение двух различных моделей агрессивного поведения, а именно модели, рассматривающей отдельные агрессивные действия, и модели, рассматривающей взаимодействие агрессора и жертвы. При этом, для каждой модели разработаны классификации со своими основаниями. Классификация, описывающая отдельные агрессивные действия, имеет следующие основания: агрессивность действий, тип сопутствующего поведения, модальность проявления агрессии, направленность агрессивного действия. Классификация, описывающая взаимодействие агрессора и жертвы расширяет предыдущую и имеет дополнительные основания: участник взаимодействия, наличие ответного действия. Разработанные классификации могут быть использованы в качестве основы для разработки онтологических моделей агрессии, отражающих существенные признаки и отношения предметной области, выработки методики оценки агрессивного поведения.

2. Разработаны онтологические модели агрессивного поведения пользователей информационного пространства. Разработанные онтологии реализуют модель Resource Description Framework (RDF) и представляют собой семантические сети, которые отражают классы объектов предметной области и отношения между ними. Онтологические модели основаны на разработанных классификациях агрессии и отражают две ситуации: случай отдельных агрессивных действий и случай взаимодействия агрессора и жертвы. На основании разработанных классификаций были выделены наиболее значимые классы предметной области, а также иерархические отношения наследования признаков этих классов. Отличительной особенностью разработанных онтологических моделей является

введение специальных функциональных отношений, отражающих факт выполнения агрессивного действия, а также, объект, на который это действие направлено. Онтологическая модель, рассматривающая проявления агрессии, позволяет классифицировать физическую и вербальную агрессию, направленную на себя, неодушевленные предметы, другие одушевленные предметы безотносительно поведения жертвы. Онтологическая модель, рассматривающая взаимодействие агрессора и жертвы, расширяет предыдущую и вводит дополнительные классы и отношения, позволяющие рассматривать ответные агрессивные действия жертвы, направленные на агрессора. Разработанные онтологические модели могут служить основой для методики оценки агрессивного поведения пользователей информационного пространства, а также основой для сбора и оценки набора видеоданных для обучения нейросетевой модели распознавания агрессивного поведения.

3. Разработана методика оценки агрессивного поведения, представленного на видеоматериалах, пользователей информационного пространства в различных ситуациях. Методика основывается на общем методе не включенного наблюдения за поведением, который подразумевает оценку поведения по определенным критериям на определенных временных интервалах. Критерии оценки поведения выведены из разработанных классификаций и онтологических моделей. Отличительной особенностью методики является выведение перечня возможных агрессивных действий, относящихся к различным видам агрессии, которые представлены в разработанных классификациях и онтологических моделях. Выведение основных действий выполнялось на основе статистического анализа эмпирического материала, полученного в ходе предварительного поискового наблюдения за агрессивным поведением, представленным на видео в сети Интернет. Другой отличительной особенностью методики является выделение специфики агрессивного поведения русскоязычных пользователей. Эта специфика заключается в вербальном речевом поведении, для которого в русском языке существуют отличительные обороты. Еще одной отличительной особенностью разработанной методики является учет ответных агрессивных действий. На основании разработанной методики составлен протокол наблюдения, который может использоваться для оценки и разметки видеоданных, содержащих агрессивное поведение русскоязычных пользователей. Разработанная методика апробирована на видео в открытом доступе в сети Интернет, содержащих агрессию пользователей информационного пространства.

4. На основании разработанной методики был собран и размечен набор видеоданных, содержащий агрессивное поведение пользователей в информационном

пространстве, а именно – поведение пользователей в ходе прямых трансляций посредством сети Интернет, что обеспечивает натуральность проявляемого поведения. Отличительными особенностями этого набора данных или выборки является то, что она содержит поведение русскоязычных пользователей, отражает поведение с точки зрения взаимодействия агрессора и жертвы, а также отражает различные виды агрессии по модальности и направленности агрессивных действий. Разметка выполнялась пятью аннотаторами. Полученные в ходе разметки высокие показатели согласованности оценок (значение коэффициента Криппендорфа в среднем по выборке оставило 0.662) свидетельствуют о высокой надежности используемой методики оценки. Итоговая выборка содержит в себе экземпляры поведения агрессора и жертвы, относящиеся к вербальной и физической агрессии, которая может направлена на неодушевленный предмет, на другой одушевленный предмет, на самого агрессора (аутоагрессия). Результирующая выборка содержит 3269 экземпляров.

5. В рамках настоящей работы было выполнено обучение и тестирование нейросетевой модели многомодального распознавания агрессивных действий. В данной нейросетевой модели обрабатываются следующие модальности: двигательная активность тела на видео, мимические проявления на видео, невербальное речевое поведение на аудио, вербальное речевое поведение в транскрибированном тексте. Главное отличие данной нейросетевой модели заключается в новом способе связывания векторов признаков отдельных модальностей для формирования единого признакового пространства, которое инвариантно виду обрабатываемой модальности и зависит только от распознаваемого класса. Инвариантность обеспечивается посредством обработки векторов признаков отдельных модальностей одной и той же нейронной сетью, которая в ходе обучения настраивает свои параметры таким образом, чтобы результат не зависел от вида модальности, а зависел только от распознаваемого класса. Другими отличительными особенностями этой модели являются: возможность работы в условиях недостатка или отсутствия информации отдельных модальностей; возможность выполнять распознавание различных видов агрессии, представленных в выборке. Обучение и тестирование модели выполнялось на выборке многомодальных данных, описанных в предыдущем пункте. Показанные в ходе тестирования оценки результатов распознавания (75.2% точности для вербальной агрессии и 73.1% точности для физической) свидетельствуют о высокой эффективности применения многомодального подхода обработки данных нейросетевой моделью по сравнению с обработкой одной модальности (67.6% точности для вербальной агрессии и 61.5% точности для физической агрессии).

Степень новизны полученных результатов

Вопросам составления и оценки наборов данных, содержащих проявления агрессивного поведения посвящен ряд работ зарубежных исследователей. В ряде работ [Lefter I., Rothkrantz L.J.M., Burghouts G.J. A comparative study on automatic audio–visual fusion for aggression detection using meta-information // Pattern Recognition Letters.- 2013. – P. 1953-1963.][Lefter I. et al. Addressing multimodality in overt aggression detection //International Conference on Text, Speech and Dialogue. – Springer, Berlin, Heidelberg, 2011. – С. 25-32.][Yang Z. Multi-modal aggression detection in trains. – 2009.][Lefter I., Burghouts G. J., Rothkrantz L. J. M. Learning the fusion of audio and video aggression assessment by meta-information from human annotations //2012 15th International Conference on Information Fusion. – IEEE, 2012. – С. 1527-1533.] [Lefter I. Multimodal Surveillance: Behavior analysis for recognizing stress and aggression. – 2014.] исследовательской группы Технологического университета Делфта (Нидерланды) были сформированы наборы многомодальных данных, содержащих агрессивное поведение людей в пассажирских вагонах и на станциях железных дорог Нидерландов. Агрессивное поведение исполнялось актерами согласно сценариям агрессивного поведения. Сами сценарии были сформированы на основании онтологической модели агрессии в железнодорожном транспорте, полученной на основании экспертных мнений работников службы безопасности железных дорог. Не натуральность проявлений агрессивного поведения снижает экологическую валидность данных наборов. Однако, сформулированные в ходе исследования критерии оценки агрессивного поведения на железной дороге повышают ценность данного исследования. Также стоит отметить малый объем выборки, что не позволяет использовать современные глубокие нейронные сети для распознавания агрессивного поведения на основании этих выборок.

В работах [Lefter I., Burghouts G. J., Rothkrantz L. J. M. An audio-visual dataset of human–human interactions in stressful situations //Journal on Multimodal User Interfaces. – 2014. – Т. 8. – №. 1. – С. 29-41.] той же группой исследователей рассматривается составление набора данных, содержащего агрессивное поведение в различных локациях железнодорожного вокзала: на стойке информации. Анализируются различные аспекты поведения людей в данных локациях. Агрессивное поведение в данных наборах исполнялось актерами на основании сценарных заготовок, задающих рамки взаимодействия в рассматриваемых ситуациях: медленное обслуживание спешащего человека, ошибочные консультации, выдаваемые сотрудниками службы информации, прекращение обслуживания клиентов в связи с перерывом на обед, сотрудник службы информации разговаривает по телефону и не обращает внимание на клиентов. Сами взаимодействия представляют собой импровизацию. Исследование концентрируется на

проявляемых жестах рук и тела в совокупности с вербальным и невербальным речевым поведением. Приводится список исследуемых жестов, основанный на анализе известной литературы. Однако, другие признаки, которые могут быть связаны с агрессивным поведением, не рассматриваются. Также следует выделить малый объем выборки с точки зрения представленных в наборе людей.

Работа [Lefter I., Rothkrantz L. J. M. Multimodal cross-context recognition of negative interactions //2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW). – IEEE, 2017. – С. 56-61.] расширяет описанные выше ситуации, которые могут произойти у стойки информации, и включает в себя также проявления агрессивного поведения около камер хранения, в кафе и около торговых автоматов. При этом, для каждой локации рассматриваются свои сценарные заготовки, которые также отыгрываются актерами путем импровизации. Разметка набора данных выполнялась с точки зрения проявляемого участниками уровня стресса, а также уровня негативности ситуации. При этом эти две оценки выполнялись для всей оцениваемой ситуации без выделения отдельных участников. Также, авторы не рассматривали анализ отдельных проявлений вербального и невербального поведения, которые характерны для проявлений агрессивного поведения. В дополнение стоит сказать, что набор данных имеет сравнительно малую выборку с точки зрения представленных в нем людей.

Авторами работы [Lefter I. et al. NAA: A multimodal database of negative affect and aggression //2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII). – IEEE, 2017. – С. 21-27.] представлен набор данных, содержащий взаимодействие двух людей в стрессовых ситуациях. Рассматриваются три сценарных заготовки, обыгрывающих обыденные ситуации, в которых может проявляться вербальная агрессия: опаздывающий на важное собеседование пассажир не может заплатить за проезд вследствие отсутствия денег, человек не может получить верхнюю одежду в гардеробе из-за утерянного номерка, регистрация студента на образовательный курс после окончания дат приема. Само взаимодействие представляет собой импровизацию. При этом, проявляющий агрессию участник взаимодействия является актером, который способен высоко выразительно симитировать агрессивное поведение, а тот участник, на кого направлено это поведение, актером не является, и ему необходимо адекватно реагировать на ситуацию. Данный набор содержит видеозаписи вербального взаимодействия двух людей, один из которых проявляет агрессивное поведение. Помимо кадров видео с записанным поведением, аудиодорожки, содержащей невербальное речевое поведение, также присутствует транскрипция реплик людей и записи психофизиологических показателей

(кожно-гальваническая реакция, электрокардиограмма, электромиограмма). Полученные в результате отдельные реплики оценивались независимыми аннотаторами по следующим параметрам: уровень агрессии, уровень страха, знак, интенсивность, доминантность (valence, arousal, dominance). При этом, авторами не выделялись отдельные действия, которые специфичны для проявления агрессии. Кроме того, необходимо отметить малый объем выборки с точки зрения представленных на видео людей и рассматриваемых ситуаций. Также приводится также статистическое исследование этого набора данных с точки зрения поиска связей между оцененными параметрами и психофизиологическими показателями [Lefter I., Fitrianie S. The multimodal dataset of negative affect and aggression: a validation study //Proceedings of the 20th ACM International Conference on Multimodal Interaction. – 2018. – С. 376-383.].

В работе [Nievas E. B. et al. Violence detection in video using computer vision techniques //International conference on Computer analysis of images and patterns. – Springer, Berlin, Heidelberg, 2011. – С. 332-339.] приведены два набора видеоданных. Первый набор содержит видеозаписи драк между хоккеистами в ходе матчей Национальной хоккейной лиги. Второй набор содержит видеозаписей драк в различных фильмах. Данные наборы данных имеют следующие особенности: упор делается на одной модальности – кадрам видео, содержащим поведенческие проявления, при этом, аудиодорожки у многих видео отсутствуют; во-вторых, малая вариативность проявляемого поведения, что весьма затрудняет его анализ и обобщение за пределы рассматриваемых ситуаций (хоккейных и кинематографических драк); малый объем выборок данных.

Авторами работы [Hassner T., Itcher Y., Kliper-Gross O. Violent flows: Real-time detection of violent crowd behavior //2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. – IEEE, 2012. – С. 1-6.] составлен набор данных содержащий проявления агрессивного поведения в толпе – драки, стычки с полицией, потасовки и т.д. У данного набора представлен не однородный состав проявлений агрессивного поведения в толпе: в наборе представлены различные ситуации – футбольные матчи, демонстрации, транспорт, однако, в каждой из этих ситуаций репрезентативно не представлены примеры агрессивного поведения. Кроме того, авторами не приводятся критерии отбора, отличающие агрессивное поведение толпе от не агрессивного. В данном наборе у видеозаписей отсутствуют аудиодорожки, что делает невозможным многомодальный анализ агрессивного поведения. Также стоит отметить малый объем выборки данных.

В работе [Perez M., Kot A. C., Rocha A. Detection of real-world fights in surveillance videos //ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – IEEE, 2019. – С. 2662-2666.] приведено описание набора данных CCTV-Fights, содержащего записи камер внешнего видеонаблюдения, на которых присутствует агрессивное поведение двух и более людей. Видео для этого набора были собраны на популярных видеохостингах, таких как YouTube. Не смотря на однородность ситуаций, в которых проявляется агрессивное поведение, а также на сравнительно большой объем выборки, данный набор данных имеет следующие недостатки: во-первых, у многих видеозаписей отсутствует аудиодорожка, а у других оригинальная аудиодорожка заменена на музыкальное сопровождение; во-вторых, нет разделения на проявления агрессивного поведения между двумя людьми и между группой людей; в-третьих, не приводятся критерии, позволяющие отличить агрессивное поведение от не агрессивного.

В работе [Sultani W., Chen C., Shah M. Real-world anomaly detection in surveillance videos //Proceedings of the IEEE conference on computer vision and pattern recognition. – 2018. – С. 6479-6488.] приводится большая выборка видео, на которых содержатся различные аномальные ситуации, которые могут быть напрямую не связаны с агрессивным поведением. Однако, в этом наборе также присутствует большая выборка видео, на которых содержатся проявления агрессивного поведения: нападения, поджоги, ограбления, кражи, стрельба, вандализм. Главной отличительной особенностью этого набора данных является его слабая разметка – не приводятся конкретных временных рамок, ограничивающих проявление того или иного поведения на видео. Кроме того, авторами не приводятся критерии, по которым следует отличать один вид поведения от другого.

Авторами работы [Cheng M., Cai K., Li M. Rwf-2000: An open large scale video database for violence detection //2020 25th International Conference on Pattern Recognition (ICPR). – IEEE, 2021. – С. 4183-4190.] также приведено описание набора данных, содержащего записи камер внешнего видеонаблюдения, на которых присутствует агрессивное поведение двух и более людей. Для этого набора характерны те же самые недостатки, что и для набора CCTV-Fights. Отличительной особенностью этого набора является более равномерное временное разбиение видео на фрагменты, а также больший объем выборки.

В результате анализа известных подходов показал недостаточную проработанность вопросов составления формализованных представлений предметной области человеческой агрессии и выработки на основании этих представлений обоснованных критериев оценки с целью формирования обучающих выборок для систем машинного обучения. В частности,

можно сделать ряд важных выводов: во-первых, недостаточно проработаны вопросы выделения отдельных поведенческих актов или сочетаний таких актов, которые могут быть использованы в качестве признаков или критериев оценки, и по которым следует определять агрессивное поведение в ходе сбора данных и их разметки. Во-вторых, для оценки агрессивного поведения часто используется обыденное представление об агрессии, агрессия зачастую путается с гневом, демонстративным поведением, не рассматриваются методики, направленные на выделение таких признаков. В основе формализаций, например, посредством онтологий, предметной области агрессивного поведения также лежат недостаточно обоснованные критерии. В-третьих, часто не анализируется специфика и контекст ситуаций, в которых проявляется агрессия, явным образом не выделяются различия в проявлениях агрессии в различных ситуациях. В-четвертых, не рассматривается взаимодействие агрессора и жертвы агрессии. В-пятых, отсутствуют наборы данных, содержащие проявления агрессии русскоязычных пользователей на видео. В-шестых, известные наборы данных страдают от низкой экологической валидности, которая заключается в том, что многие наборы данных содержат не натуральные проявления поведения, а сыгранные актерами по сценарию. В-седьмых, в известных наборах зачастую не достаточные объемы выборок с точки зрения количества людей, представленных в этих наборах. В-восьмых, в известных работах недостаточно внимания уделяется этапу предварительного составления перечня действий, входящих в состав изучаемого поведения, и выделения наиболее часто встречающихся действий. Данная процедура основана на методике внешнего поискового наблюдения и является обязательной в психологических, антропологических и социологических исследованиях, посвященных изучению поведения человека [Wieder G. B., Weiss R. L. Generalizability theory and the coding of marital interactions //Journal of Consulting and Clinical Psychology. – 1980. – Т. 48. – №. 4. – С. 469.][Yudofsky S. C. et al. The Overt Aggression Scale for the objective rating of verbal and physical aggression //The American journal of psychiatry. – 1986.][Jacob T. et al. Family Interaction in the Home Development of a New Coding System //Behavior Modification. – 1995. – Т. 19. – №. 2. – С. 147-169.][Валявко С. М., Князева А. А. Изучение личностного развития дошкольников с помощью стандартизированного наблюдения //Системная психология и социология. – 2015. – №. 2. – С. 15-24.]. Составленный при помощи этой процедуры перечень действий в дальнейшем может быть использован в качестве критериев оценки изучаемого вида поведения. Это, в свою очередь, также может быть использовано при отборе и разметке данных для обучения моделей машинного обучения. В ходе составления и разметки наборов данных будут использованы обоснованные критерии оценки, а не обыденное представление об изучаемом поведении, что, в свою очередь, повысит качество

составляемых наборов поведенческих данных. В связи с этим, а также со всеми перечисленными недостатками, встает актуальный вопрос о более глубокой проработке классификаций, формализованных представлений, а также методик составления и разметки наборов данных, содержащих агрессивное поведение, особенно в контексте оценки поведения русскоязычных пользователей.

Методы и подходы, использованные при реализации проекта

Разработка классификации видов агрессии, проявляемой, в информационном пространстве

Важным условием успешной реализации проекта являлась необходимость произведения выработки обоснованного представления предметной области агрессии, проявляемой в различных ситуациях. Такое представление должно содержать существенные признаки предметной области, а также существенные связи и отношения между этими признаками, т.е. наиболее полно отражать описываемую предметную область. Предлагаемые классификации способствуют выработке обоснованных критериев оценки поведения для сбора и разметки валидных выборок данных, которые будут использоваться для обучения моделей машинного обучения.

Отправной точкой разработки классификации является выбор базового определения агрессии, из которого будут определяться ключевые признаки агрессивного поведения пользователей, лежащие в основу разрабатываемой классификации. Также необходимо упомянуть ряд других важных классификаций, которые можно принять для основы разрабатываемой в ходе исследования.

В качестве базового было взято определение, данное отечественным исследователем С.Н. Ениколоповым: «Агрессия суть целенаправленное деструктивное и наступательное поведение, нарушающее нормы и правила сосуществования людей в обществе, наносящее вред объектам нападения (одушевленным и неодушевленным), причиняющее ущерб или вызывающее у них психологический дискомфорт: отрицательные переживания, состояния напряженности, страха, подавленности и др.» [Ениколопов С. Н. Понятие агрессии в современной психологии //Прикладная психология. – 2001. – №. 1. – С. 60-72.]. В этом определении особо следует выделить следующие ключевые признаки, на основании которых в данной работе строится классификация агрессии:

1. Деструктивный характер действия.
2. Целенаправленность действия.

3. У агрессивного действия есть объект актор или субъект действия.
4. У агрессивного действия есть свой объект или жертва.

Одной из самых распространенных классификаций агрессивного поведения является классификация, выведенная зарубежным исследователем А. Бассом [Buss A. H. The psychology of aggression. – Wiley, 1961.]. Она имеет следующие основания:

1. Модальность проявления агрессии: Физическая – вербальная.
2. Активность проявления агрессии: Активная – пассивная;
3. Открытость действий: Прямая – косвенная.

Прежде, чем вывести основания для разрабатываемой классификации агрессивного поведения пользователей социкиберфизического пространства, представленного на видео, на основании приведенных выше определений и классификаций, необходимо выделить общие особенности поведения пользователей социкиберфизического пространства.

Сначала следует ограничить множество ситуаций, которые рассматриваются в данном исследовании. В качестве одной из наиболее распространенных ситуаций, встречающихся в социкиберфизическом пространстве, в которых проявляется агрессивное поведение пользователей, являются прямые трансляции или видеостримы, организованные посредством сети Интернет. Такие трансляции проводятся с различными целями: научно-просветительские, развлекательные, выполняющие функции СМИ и т.д. Описывая происходящее на этих трансляциях, необходимо выделить те общие черты, присущие большинству таких трансляций. Обычно на кадрах видеотрансляции присутствует один или несколько участников, которые взаимодействуют между собой и со зрителями стрима посредством дополнительных аудио, видео и текстовых средств коммуникации. Трансляции ведутся в подавляющем большинстве случаев из помещений, имеющих различный интерьер (жилые, офисные, помещения, специально оборудованные студии, общественные пространства и т.д.) Помимо этого, трансляции могут вестись из открытых уличных пространств. Сами участники обычно располагаются сидя перед камерой, зачастую, за столом. Также возможно расположения участников перед камерой стоя в полный рост или лежа. Стоит особо отметить, что важной составляющей взаимодействия участников трансляции и её аудитории являются денежные пожертвования, направляемые зрителями трансляции её участникам. Обычно пожертвование сопровождается вопросом, пожеланием, какой-либо иной репликой или, что более важно, исполнением участниками трансляции какого-либо действия, желаемого жертвующим деньги зрителем. При этом, реплики, сопровождаемые пожертвованиями, в подавляющем большинстве случаев имеют приоритет во внимании участников трансляции.

Таким образом обеспечивается мотивация оживлённого общения участников стрима со своей аудиторией. Однако, именно в этой механике взаимодействия может лежать потенциальная угроза мотивированных и поощряемых проявлений агрессии одних участников трансляции по отношению к другим.

В данном исследовании наибольший интерес представляют видеотрансляции, именуемые в СМИ и литературе "треш-стримами". Актуальность рассмотрения именно этой категории обусловлено, во-первых, их относительной распространённостью на видеохостинговых платформах в сети Интернет, а во-вторых, в последнее время произошло несколько резонансных событий, исходом которых было нанесение вреда здоровью или даже гибель участников треш-стримов [<https://www.bbc.com/russian/features-55238100>]. Всё это особо остро ставит проблему выявления агрессивного поведения участников треш-стримов с целью оперативного реагирования на такое поведение и недопущения пагубных последствий.

Треш-стримы являются трансляциями развлекательного характера, в ходе которых обычно происходят эксцессы различного характера, включая агрессивное поведение. Помимо не явных и трудно выявляемых мотивационных факторов агрессивного поведения (например, личностные особенности, внутренние установки участников и взаимодействующих с ними зрителей), существуют также и очевидные факторы, которые если напрямую и не поощряют агрессивное поведение, то, по крайней мере не предотвращают его.

К факторам, которые напрямую поощряют агрессивные действия, в первую очередь, относится выставление организаторами трансляций прямых преискурентов, которые явным образом определяют денежную стоимость определенного агрессивного действия. К менее явным факторам, способствующим агрессивному поведению, необходимо отнести недостаточность или полное отсутствие модерации оскорбительных или иных реплик, которые могут спровоцировать агрессию, в ходе взаимодействия аудитории и участников стрима.

Выделение оснований классификации

В первую очередь, необходимо выделить отличительные признаки, которые характеризуют агрессивное поведение пользователей в рассмотренной ситуации треш-стримов. Сначала необходимо выделить виды агрессии, которые с наибольшей вероятностью будут встречаться в рассматриваемых ситуациях.

Очевидно, что наиболее распространенными видами агрессии будут активная прямая физическая и активная прямая явная вербальная виды агрессии. Эти виды агрессии отличаются, в первую очередь, тем, что могут быть выражены явным образом посредством множества отдельных актов или действий, которые можно однозначно классифицировать как агрессивные (например, удары, поломка предметов для физической агрессии или оскорбления, угрозы для вербальной агрессии). Что касается пассивной и косвенной агрессии, то данные виды агрессивного поведения, во-первых, имеют не достаточную определенность, нечеткость и размытость формулировок действий, входящих в их состав, во-вторых, действия, входящие в состав косвенной и пассивной агрессии достаточно сложно внешне явным образом наблюдать, сами эти действия проблематично локализовать в определенных, четко очерченных временных интервалах. Вследствие всего вышеперечисленного, эти виды агрессии недостаточно распространены в видео, содержащего агрессивное поведение на треш-стримах, в открытом доступе в сети Интернет, и рассматриваться нами в разрабатываемой классификации не будут.

Кроме того, данное выше определение агрессии подразумевает объект нападения или жертву агрессии. В этом случае агрессия подразделяется на агрессию на неодушевленные предметы, агрессию, направленную на одушевленные предметы, и агрессию, направленную на самого себя или аутоагрессию.

Несмотря на то, что большинство исследователей сходятся на мнении, что эмоциональные состояния не являются ни необходимыми, ни достаточными критериями для классификации агрессии, эти эмоциональные состояния, а также дополнительные поведенческие акты (мимические выражения, двигательная активность тела, жестикация, невербальное речевое поведение) очень часто сопутствуют агрессивному поведению. Это делает их выделение и анализ особенно интересным не только с точки зрения составления автоматических систем распознавания агрессивного поведения, но, что более важно, и с точки зрения анализа взаимосвязи агрессивных действий и сопутствующих эмоционально-поведенческих проявлений.

Поэтому, с учетом всего вышеперечисленного, в качестве оснований классификации были использованы следующие критерии:

1. Модальность: вербальная – физическая.
2. Направление: на другой одушевленный предмет, на не одушевленный предмет, на самого себя (аутоагрессия).
3. Сопутствующие эмоциональные и поведенческие проявления, которые нельзя напрямую отнести к агрессивному поведению. В качестве конкретных

эмоционально-поведенческих проявлений в данной работе выступают следующие:

- Мимические выражения;
- Двигательная активность корпуса тела и головы;
- Активная жестикуляция;
- Выраженное невербальное речевое поведение (высота и громкость голоса).

Схема, отражающая указанные выше критерии классификации и их соотношение, приведена на рисунке 1.



Рисунок 1. Классификация агрессии с точки зрения отдельных агрессивных действий

Важно отметить, что ввиду недостаточной разработанности методик количественной оценки интенсивности проявлений поведения русскоязычных пользователей на видео, все обозначенные виды представлены дихотомическими переменными, отражающими наличие или отсутствие интересующего нас поведения на оцениваемом участке видео. Также необходимо особо подчеркнуть, что в данной работе по очевидным причинам не рассматривается вербальная агрессия на неодушевленные предметы: неодушевленному предмету невозможно нанести вред словесными оскорблениями, проклятиями, угрозами и прочими агрессивными актами вербальной агрессии. Еще одним важным замечанием будет то, что приведенные виды агрессии не могут полностью исключить друг друга (например, человек может сопровождать физическую агрессию вербальной).

Модель поведения агрессор – жертва

Приведенная выше классификация позволяет определять отдельные агрессивные действия, проявляемые пользователями информационного пространства. Однако, она не учитывает взаимодействие двух участников акта агрессии – субъекта или агрессора и объекта или жертвы. Приведенная выше классификация в полной мере отражает только те агрессивные акты, в которых в качестве объекта агрессии выступает неодушевленный предмет, либо те ситуации, в которых одушевленный объект агрессии ведет себя пассивно и не отвечает на деструктивные действия агрессора. Однако, зачастую, это не так. Нередки случаи, когда в ответ на вербальную и/или физическую агрессию по отношению к себе, жертва проявляет ответную агрессию, которая либо следует за действиями агрессора (обычно это характерно для вербальной агрессии, где агрессивные реплики чередуют друг друга), либо выполняется одновременно (это больше характерно для физической агрессии, когда два участника агрессивного акта одновременно наносят друг другу повреждения). Стоит заметить, что в последнем случае достаточно проблематично выделить агрессора и жертву: т.к. действия выполняются одновременно, оба участника одновременно являются и агрессорами, и жертвами. Несмотря на сложность выделения агрессора и жертвы, действия, входящие в такую структуру взаимодействия все же можно с высокой степенью уверенности отнести к агрессивным.

Рассмотрим подробнее сценарии, которые могут разворачиваться в рамках этой модели:

1. Агрессор атакует, жертва не отвечает.
2. Агрессор сначала атакует, а затем жертва отвечает

3. Невозможно выделить агрессора и жертву: оба участника нападают друг на друга одновременно

Второй третий сценарии существенно отличаются от первого наличием ответных агрессивных действий со стороны жертвы агрессии. При этом, между собой они отличаются только временной локализацией ответных агрессивных действий жертвы. Это не является принципиально важным для определения взаимной агрессии. Поэтому, в качестве дополнительного критерия, расширяющего предоставленную выше классификацию на взаимодействие агрессора и жертвы, достаточно принять сам факт ответных агрессивных действий.

Таким образом, в приведенную выше классификацию добавляется бинарный критерий наличия ответной агрессии, отражающий факт взаимодействия агрессора и жертвы. В случае, если объект агрессии является одушевленным, для него необходимо определить, отвечает ли он агрессией на агрессию. В случае утвердительного ответа рассматриваются виды агрессии по модальности – вербальная и физическая, а затем рассматривается направленность агрессии. При этом рассматривается только один вид агрессии – на другие одушевленные предметы. Другие направления агрессии не рассматриваются ввиду того, что они исчерпывающе покрываются предыдущей классификацией. Схема, отражающая указанные выше критерии классификации и их соотношение, приведена на рисунке 2.

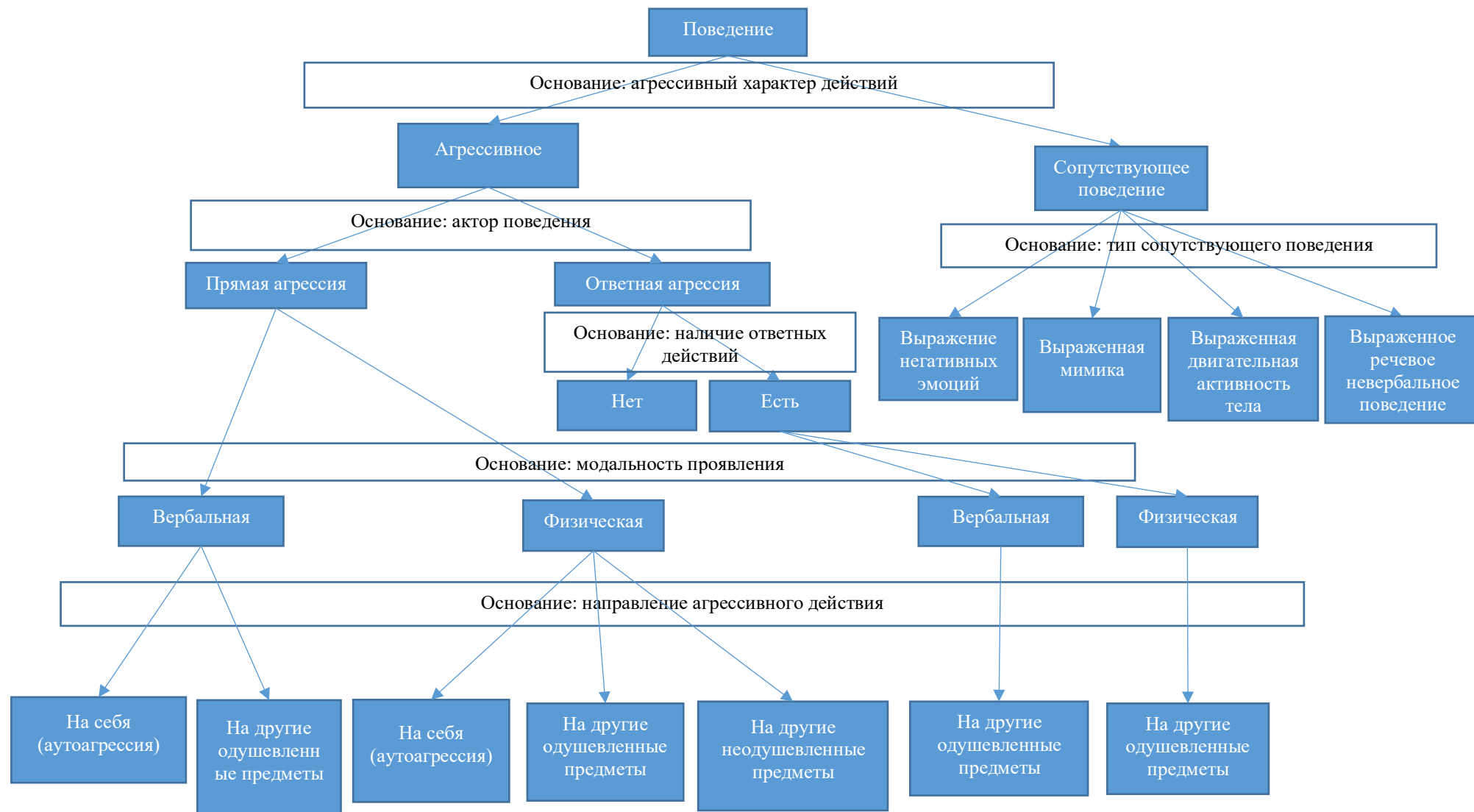


Рисунок 2. Классификация агрессии в рамках взаимодействия агрессора и жертвы

Приведенные в настоящем исследовании классификации в полной мере отражают наиболее распространенные единичные акты агрессии, а также акты взаимодействия агрессора и жертвы, когда жертва отвечает агрессией на агрессию. Разработанные классификации можно использовать для построения более детальных онтологических моделей агрессивного поведения и методик оценки агрессивного поведения, которые могут быть использованы для разработки систем машинного обучения, выполняющих автоматическое распознавание агрессивного поведения, представленного на видео, пользователей социокиберфизического пространства.

Разработка онтологических моделей агрессивного поведения пользователей в информационном пространстве в различных предметных ситуациях

Представленные выше классификации позволяют определить понятие агрессии и выделить в нем наиболее значимые признаки. Приведенные классификации отражают отношения наследования признаков частных понятий от более общих. Однако, эти классификации не отражают дополнительных свойств и отношений между понятиями. Таким образом, приведенные классификации не являются достаточными для наиболее полной формализации предметной области агрессивного поведения пользователей, социокиберфизического пространства, представленного на видео.

Для решения этой задачи наилучшим подходом является построение онтологических моделей, которые как раз и будут выражать наиболее полный перечень классов объектов, их признаков, а также отношений между классами и признаками. Онтологические модели имеют ряд неоспоримых преимуществ, к которым, в первую очередь, следует отнести следующие: для построения онтологической модели необходимо выполнение детального анализа предметной области, результатом которого является явное определение классов объектов и явлений, входящих в предметную область, их свойств и отношений между ними; возможность выполнять классификацию объектов экземпляров классов предметной области посредством продукционных правил вывода, включенных в онтологическую модель; явная интерпретируемость результатов вывода, включая промежуточные этапы.

Помимо этого, онтологические модели можно применять также и не явным образом. Например, для построения на их основе систем нечеткого вывода, вероятностных графовых моделей и прочих систем машинного обучения. Кроме того, онтологическое описание предметной области позволяют вывести правила, которые могут стать основой для методик составления и разметки обучающих выборок сложных гетерогенных данных: описание

классов, признаков и отношений явным образом можно трансформировать в инструкции для аннотаторов, которые будут в дальнейшем оценивать каждый отдельный экземпляр данных для того, чтобы пометить его тем или иным классом, входящим в онтологию. Полнота описания предметной области онтологической моделью и явное определение всех необходимых для классов рассматриваемой предметной области должны обеспечивать, в первую очередь, высокую надежность или воспроизводимость результата оценки. Таким образом, обеспечивается высокое качество оценки данных и, как следствие, составление репрезентативных обучающих выборок, которые наиболее полным образом отражают предметную область и на основе которых затем возможно производить обучение различных систем машинного обучения.

Семантическая сеть

Одним из самых распространенных и наиболее удобных средств реализации онтологических моделей являются семантические сети. Семантические сети являются графовыми моделями представления знаний. Это ориентированные графы, в которых вершинами графа являются классы или понятия, описываемые в предметной области, а ребрами являются отношения между этими понятиями или классами. При этом направление ребра соответствует направленности отношения на класс. В самом общем случае, и понятия, и отношения между этими понятиями зависят от конкретных реальных свойств объектов в описываемой предметной области. С математической точки зрения это означает, что модель семантических сетей не накладывает никаких ограничений на представления знаний, пока их можно выразить посредством ориентированного графа особой структуры, которая подробнее будет раскрыта ниже. Такая свобода позволяет задавать отношения, которые наиболее полно и точно отображают связи между классами объектов и их признаками в рассматриваемой реальной предметной области.

RDF-модель

Фреймворк описания ресурсов (Resource Description Framework – RDF) является, по сути, теоретико-множественной моделью, которая выражает отношения или связи между понятиями (классами) объектов и их свойствами или признаками посредством упорядоченных множеств или кортежей-триплетов унифицированного вида Объект-Атрибут-Значение (Object-Attribute-Value – OAV):

$$\langle O, A, V \rangle, \quad (1)$$

где O – класс объектов или отдельный объект, A – определенное свойство, связывающее класс объектов или объект с его значением; V – значение, которое атрибут принимает для класса объектов или отдельных объектов.

Стоит выделить один ключевой принцип, которому подчиняется RDF, а именно свойствоцентричность. Это означает, что при составлении семантических сетей мы можем произвольно соединять классы, входящие в онтологию, посредством заданных для этой онтологии свойств, группируя классы вокруг свойств, а не свойства вокруг классов, как это делается в объектной модели данных. Данный принцип позволяет строить представление знаний вокруг свойств классов объектов и обеспечивает свободу в организации представления знаний, что в итоге позволяет наиболее полным образом отразить все специфические свойства каждого класса объектов, включенных в онтологию.

Важным преимуществом RDF является его распространенность, стандартизация, наличие поддерживаемых сообществом языков описания и программных средств реализации. В данном исследовании для построения онтологической модели агрессивного поведения русскоязычных пользователей социкиберфизического пространства, представленного на видео, был использован Ontology Web Language (OWL) [<https://www.w3.org/OWL/>], который поддерживает принципы RDF, а также свободно распространяемая в открытом исходном коде среда Protégé [<https://protege.stanford.edu/>].

Описание классов и отношений между классами для онтологии отдельных агрессивных действий

Для того, чтобы представить онтологию в виде семантической сети необходимо сначала формализовать все классы, входящие в рассматриваемую предметную область, а также отношения между классами. Классификация, приведенная на рисунке 1, отражает иерархию классов, и, как следствие эта классификация отражает только отношения типа АКО (A Kind Of – является видом чего-то).

Рассмотрим сначала подробнее структуру исполнения акта агрессии с точки зрения актора исполнения этого действия или агрессора. На естественном языке это можно выразить как «Агрессор совершает деструктивные наступательные действия по отношению к объекту нападения». Это может быть выражено схемой, расположенной на рисунке 3.

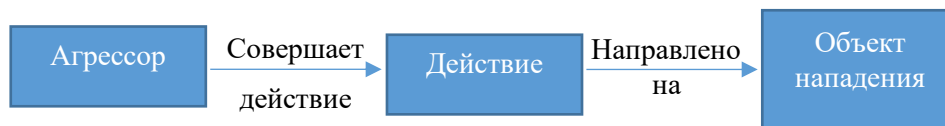


Рисунок 3. Схема агрессивного действия

Стоит отметить, что данная схема приводится в самых общих чертах, однако уже она позволяет выделить существенные связи, которые присущи каждому агрессивному акту. Эти связи или отношения дополняют АКО отношения и вместе с ним могут служить для выражения отдельного агрессивного действия. Суммируя все вышеперечисленное, в разрабатываемой онтологии отдельных агрессивных действий:

1. АКО – отношения, выражающие отношения наследования свойств более общих классов классами менее общими и более специфичными;
2. «Совершает действие» – отношение, выражающее факт исполнения действия агрессором. Данное отношение соединяет экземпляры класса «Агрессор» с экземплярами класса «Действие»;
3. «Направлено на» – отношение, выражающее направление агрессивного действия на объект нападения. Данное отношение соединяет экземпляры класса «Действие» с экземплярами класса «Объект нападения».

Для составления целостной онтологии, помимо отношений, необходимо также перечислить основные классы, которые в нее входят, включая те, которые были описаны в разделе, посвященном разработке классификации отдельных агрессивных действий. В онтологию включены следующие классы и подклассы, образующие иерархию:

– Класс «Действие»

- Класс «Агрессивное действие», который разделяется по основанию модальности проявления:
 - Класс «Вербальная агрессия», который разделяется по основанию направление агрессивного действия:
 - Класс «Вербальная агрессия на себя». К этому классу относятся такие действия как самообвинения, самооскорбления, выражение суицидальных мыслей и т.д.
 - Класс «Вербальная агрессия на другого». К этому классу относятся такие действия как матерные или не матерные оскорбления, включая оскорбления по расовому, возрастному, гендерному признакам; угрозы жизни, здоровью,

благополучию; послания; обвинения; приказы замолчать; пожелания смерти; словесное поощрение насилия;

- Класс «Физическая агрессия» который разделяется по основанию направление агрессивного действия:
 - Класс «Физическая агрессия на себя». К этому классу относятся такие действия как удары себя руками и предметами; удары головой о предметы; самопорезы; самоудушение; стрельба в себя;
 - Класс «Физическая агрессия на другого». К этому классу относятся такие действия как удары руками, ногами, головой и предметами; кидание предметов; толкание руками и ногами; борцовские приемы (захваты, подсечки, удушающие приемы, поваливание на пол); таскание за волосы; выкручивание носа; обливание жидкостью; прыжки на человека;
 - Класс «Физическая агрессия на неодушевленный предмет». К этому классу относятся такие действия как удары предметов другими предметами, руками, ногами; кидание предметов; укусы предметов; поломка предметов (разламывание предметов, отрывание от них частей и т.д.); прыжки на предметы;
- Класс «Сопутствующее действие». Это сопутствующие эмоциональные и поведенческие проявления, которые напрямую нельзя отнести к агрессии, но которые часто ее сопровождают. К ним относятся следующие: выраженное проявление эмоций; выраженная мимика; выраженная двигательная активность тела, включая жестикуляцию; выраженное невербальное речевое поведение (крики, плач и т.д.)
- Класс «Объект агрессии». Он выражает возможные объекты агрессии и разделяется по основанию одушевленности объектов на:
 - Класс «Одушевленный объект»
 - Класс «Неодушевленный объект»
- Класс «Одушевленный». Он выражает возможных участников агрессивного акта и разделяется по основанию роли агрессора или жертвы на:
 - Класс «Одушевленный объект», который выражает одушевленную жертву агрессивного действия. Данный класс разделяется по основанию, является ли агрессор сам жертвой (аутоагрессия, самоповреждение) на:

- Класс «Другой одушевленный объект»
- Класс «Сам агрессор»
- Класс «Одушевленный субъект», который выражает непосредственно агрессора. Этот класс разделяется по основанию, является ли одушевленный объект также и жертвой агрессии, на:
 - Класс «Другой одушевленный объект»
 - Класс «Сам агрессор»

Общая схема иерархии классов приведена на рисунке 4.

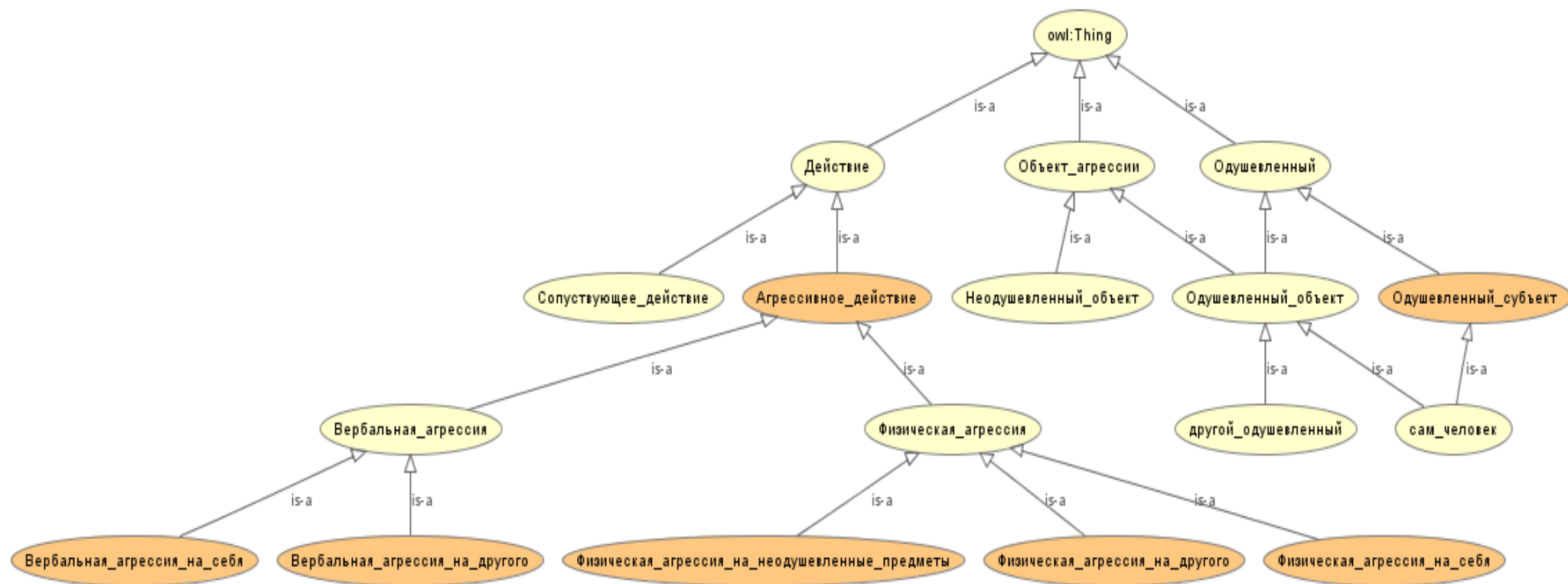


Рисунок 4. Иерархия классов онтологии отдельных агрессивных действий

Стоит особо обратить внимание на механизм множественного наследования, реализованный для классов «Одушевленный объект» и «Сам агрессор». Класс «Одушевленный объект» одновременно обладает свойствами и класса «Объект агрессии», и класса «Одушевленный». Класс «Сам агрессор», в свою очередь, является одновременным обладателем свойств и «Одушевленного субъекта», и «Одушевленного объекта». Такое совпадение агрессора и жертвы в одном лице позволяет идентифицировать проявления аутоагрессии.

Целостная онтологическая модель отдельных агрессивных действий.

Рассмотрим подробнее теперь целостную онтологию, включающую в себя все классы и отношения между ними: все классы, обозначенные на рисунке 4, а также отношения АКО, «Совершает действие» и «Направлено на». Целостная онтология в виде семантической сети приведена на рисунке 5.

Данная онтология реализует общую схему агрессивного действия, обозначенную на рисунке 3. Все подклассы класса «Одушевленный субъект» связаны с посредством отношения «Выполняет действие» со всеми действиями, включенными в класс «Действие». В случае, если действие является подклассом класса «Агрессивное действие», то оно должно быть связано с любым из подклассов класса «Объект агрессии». Эти связи имеются у каждого из следующих видов агрессивного поведения с соответствующим объектом нападения: «Физическая агрессия на неодушевленный предмет», «Физическая агрессия на себя», «Физическая агрессия на другой одушевленный предмет», «Вербальная агрессия».

Описание классов и отношений между классами для онтологической модели агрессии в рамках взаимодействия «агрессор-жертва»

Необходимо отметить, что описываемая в данном разделе онтологическая модель агрессивного поведения в рамках взаимодействия «Агрессор-Жертва» полностью включает в себя онтологическую модель, рассматривающую агрессию только с точки зрения поведения агрессора, которая была описана в предыдущем разделе. Однако, расширение классов и отношений, связывающих эти классы влечет за собой некоторые изменения в формальной организации классов и отношений в онтологии. Этим различиям и посвящен данный подраздел.

В первую очередь рассмотрим взаимодействие участников агрессивного акта в рамках модели «Агрессор-жертва». В разделе, посвященном разработке классификации агрессии в рамках взаимодействия агрессора и жертвы, были уже рассмотрены ситуации взаимодействия агрессора и жертвы, которые могут быть сведены к двум глобальным:

1. Агрессор нападает, жертва не отвечает;
2. Агрессор нападает, жертва отвечает (включая те ситуации, когда невозможно однозначно агрессора отделить от жертвы, т.к. оба участника нападают друг на друга одновременно).

На основе такого разделения можно вывести общую структуру взаимодействия агрессора и жертвы, которую можно взять за основу разрабатываемой онтологии. Эта схема приведена на рисунке 6.

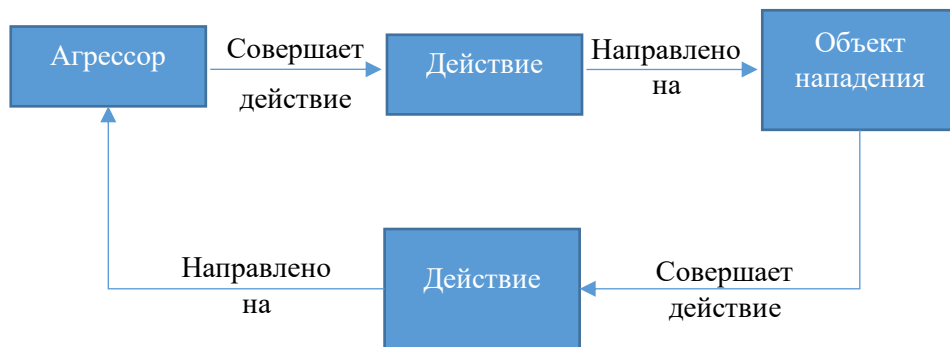


Рисунок 6. Схема взаимодействия агрессора и жертвы

Приведенная схема в самых общих чертах отражает существенные связи, которые присущи каждому агрессивному акту в рамках взаимодействия «Агрессор-жертва». Это те же самые связи или отношения, которые используются в онтологической модели, описывающей отдельные агрессивные действия. Однако, теперь они применяются также и к объекту нападения или жертве. В случае, если жертва отвечает агрессией на агрессию, она приобретает указанные свойства совершать определенные действия, направленные на агрессора.

Несмотря на то, что в предлагаемой онтологической модели используются те же самые связи, новая структура взаимодействия требует изменений в иерархии классов онтологии. Во-первых, изменения касаются класса «Агрессивные действия»:

- Класс «Агрессивное действие», который разделяется по основанию актора, исполняющего это действие на:
 - Класс «Действие агрессора», который разделяется по нескольким основаниям (объект агрессии, тип агрессии, направленность агрессии) на:
 - Класс «Агрессор: физическая агрессия», который разделяется по основанию направленности действия на:
 - Класс «Агрессор: физическая агрессия на себя»;
 - Класс «Агрессор: физическая агрессия на другого»;
 - Класс «Агрессор: физическая агрессия на неодушевленные предметы на предметы»;
 - Класс «Агрессор: вербальная агрессия», который разделяется по основанию направленности действия на:
 - Класс «Агрессор: вербальная агрессия на себя»;
 - Класс «Агрессор: вербальная агрессия на другого»;
 - Класс «Действие жертвы», который разделяется по нескольким основаниям (объект агрессии, тип агрессии, направленность агрессии) на:

- Класс «Жертва: физическая агрессия на другого»;
- Класс «Жертва: вербальная агрессия на другого»;

Во-вторых, вводятся дополнительные классы, отражающие типы агрессии:

- Класс «Агрессия», который разделяется по основанию типа или модальности проявляемой агрессии на:
 - Класс «Физическая агрессия»;
 - Класс «Вербальная агрессия».

Общая схема иерархии классов приведена на рисунке 7.

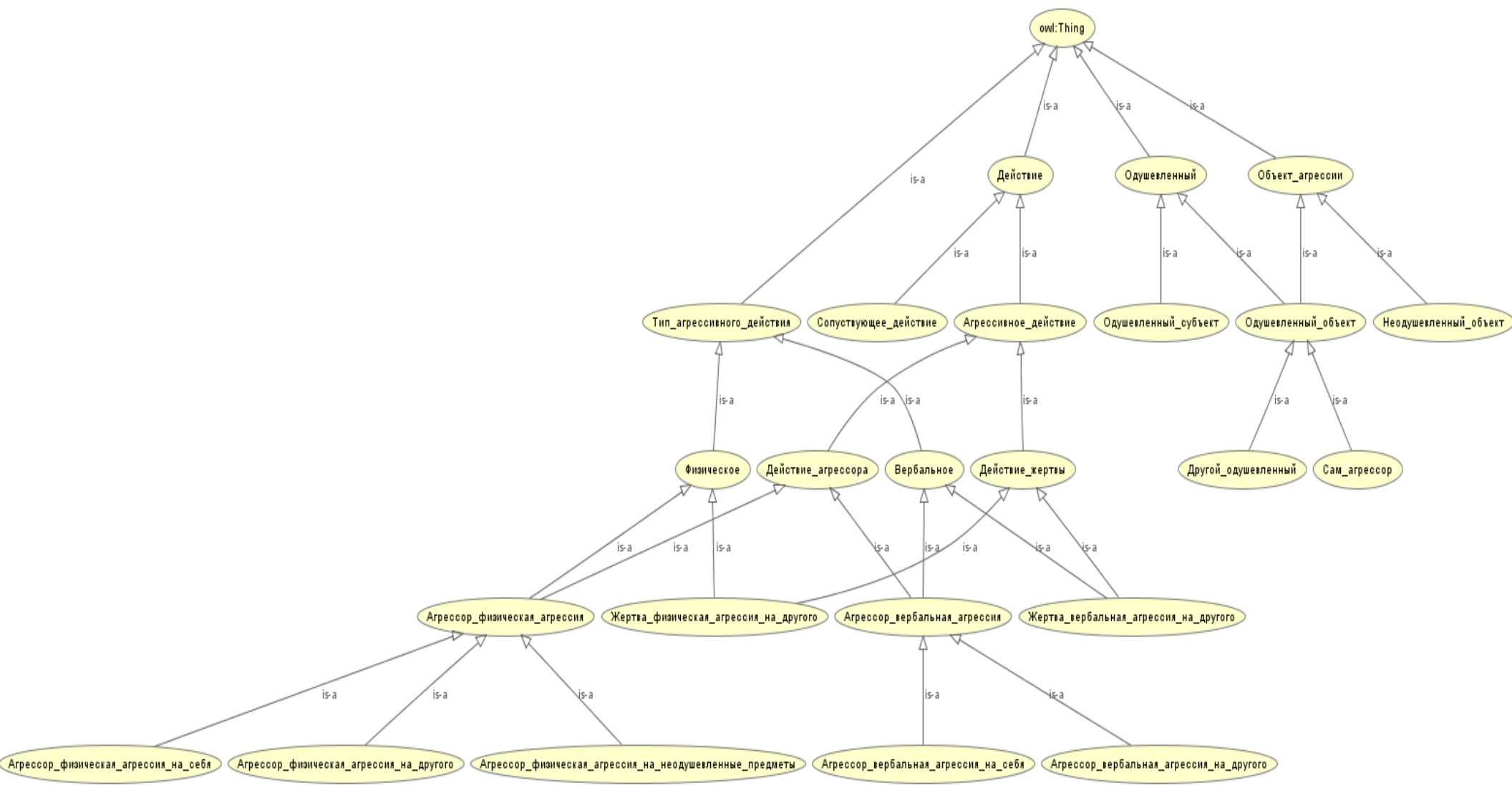


Рисунок 7. Иерархия классов онтологии агрессивного поведения в рамках модели взаимодействия «Агрессор-жертва»

Необходимо особо отметить, что в измененной иерархии классов, во-первых, отражены агрессивные и сопутствующие действия, которые выполняются одушевленным объектом нападения или жертвой. Во-вторых, вследствие того, что жертва может отвечать на действия агрессора, целесообразно выделить в отдельную иерархию классов типы агрессии (вербальная и физическая). В-третьих, в агрессивных действиях жертвы отсутствуют некоторые типы действий, которые есть у действий агрессора. Вербальная и физическая агрессия, направленная на себя, а также физическая агрессия, направленная на неодушевленные предметы, – все они отсутствуют по причине того, что эти ситуации избыточны и полностью отражены в действиях агрессора.

Целостная онтологическая модель агрессии в рамках взаимодействия «агрессор-жертва».

Рассмотрим подробнее теперь целостную онтологию, включающую в себя все классы и отношения между ними: все классы, обозначенные на рисунке 7, а также отношения АКО, «Совершает действие» и «Направлено на». Целостная онтология в виде семантической сети приведена на рисунке 8.

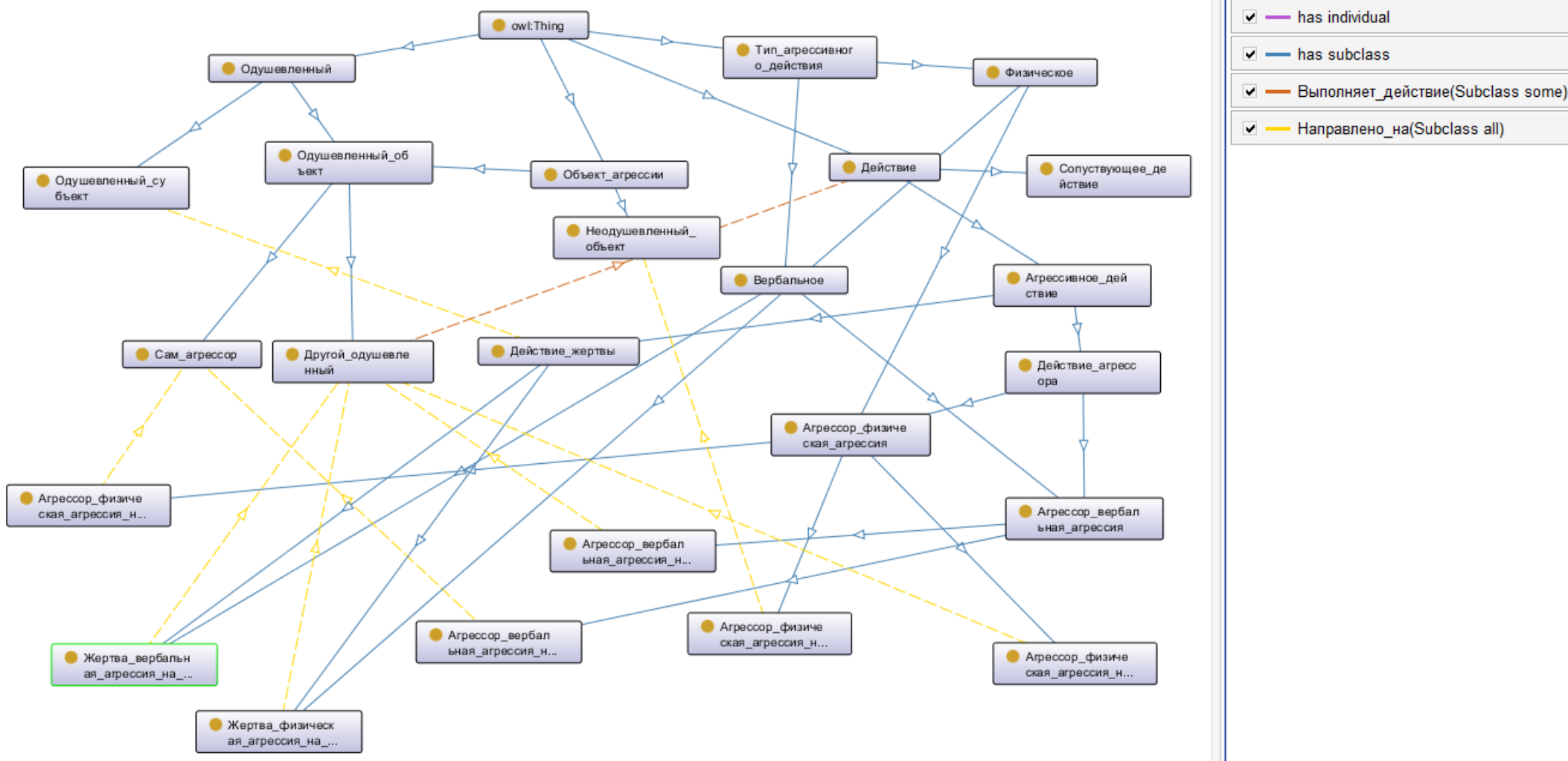


Рисунок 8. Онтология отдельных агрессивных действий

Данная онтология реализует общую схему агрессивного действия в рамках модели взаимодействия «Агрессор-жертва», обозначенную на рисунке 6. Помимо действий, которые может выполнять агрессор, данная онтологическая модель также подразумевает и ответные действия, которые может выполнять жертва. Связь актора агрессивного действия (агрессора или жертвы) с действием реализовано посредством отношения «Выполняет действие». Направленность каждого из действий также отражена в разработанной онтологии посредством отношения «Направлено на». При этом, в отличие от предыдущей онтологии, привнесена также возможность направлять агрессивное действие обратно от жертвы к агрессору.

Разработка и апробация методики оценки агрессивного поведения пользователей информационного пространства в различных предметных ситуациях

Решение задачи распознавания поведения людей на видео на данный момент лучше всего выполняют те модели, методы и системы, которые основаны на подходе нейронных сетей и машинного обучения. Однако, для эффективного функционирования данные методы требуют больших репрезентативных обучающих выборок, которые наиболее полным образом отражают свойства предметной области, в которой решаются задачи. Для составления таких выборок и оценки отдельных экземпляров в этих выборках необходима выработка релевантных методик оценки поведения пользователей социокиберфизического пространства. Эта методика должна основываться на обоснованных критериях оценки, которые наиболее полным образом отражают существенные признаки агрессивного поведения в различных ситуациях. В рамках данного исследования предлагается методика, основанная на приведенных классификациях и онтологических моделях, которые наиболее полным образом отражают существенные свойства и отношения в области агрессивного поведения пользователей социокиберфизического пространства, представленного на видео. Перейдем к более подробному описанию предлагаемой методики.

Несмотря на некоторые различия в приведенных выше онтологических моделях по структуре классов и отношений между ними, онтологическая модель, описывающая агрессивное поведение в рамках модели взаимодействия «Агрессор-Жертва», полностью включает в себя онтологическую модель, рассматривающую агрессивное поведение только с точки зрения агрессора. Вследствие этого целесообразно разрабатывать методику оценки агрессивного поведения на основе онтологической модели взаимодействия «Агрессор-Жертва».

Разрабатываемая методика полностью основывается на методе внешнего не включенного наблюдения за поведением людей, которое представлено на видео, находящихся в открытом доступе в сети Интернет. Оценка таких видео полностью исключает влияния исследователя на ситуацию, что обеспечивает экологическую валидность данных или естественность поведенческих проявлений индивидов. Также стоит отметить еще одно преимущество этого метода, а именно то, что публичная демонстрация поведения, которое может быть просмотрено неограниченным кругом лиц, снимает необходимость получать согласие на участие в наблюдении. Кроме того, внешнее наблюдение суть метод объективной оценки, потому что данные о двигательных реакциях, вербальном, невербальном поведении фиксируются извне и не зависят от личностных особенностей испытуемых.

Стоит отметить также и некоторые ограничения, которые присущи методу внешнего наблюдения, самым существенным из которых является невозможность влиять на ситуацию, изменяя стимульный материал либо условия наблюдения, что делает невозможным получение причинно-следственных связей из наблюдаемых данных (данная проблема усугубляется еще и тем, что постановка эксперимента по исследованию агрессивного поведения в принципе невозможная, т.к. противоречит этическим принципам проведения психологических экспериментов). Другим важным ограничением является невозможность измерять у испытуемых те же самые исследуемые характеристики другими методами, а также невозможность получить какие-либо дополнительные характеристики испытуемых, что существенно затрудняет валидацию методики. Но, несмотря на все указанные ограничения, именно метод внешнего наблюдения наилучшим образом подходит в качестве основы для методик разметки видеоданных, содержащих поведение людей.

Существенные признаки агрессии, лежащие в основе методики

Разработанная методика основывается на приведенном определении агрессии, из которого выводится соответствующая классификация агрессивного поведения и онтологическая модель. Еще раз напомним это определение: агрессия суть «целенаправленное деструктивное и наступательное поведение, нарушающее нормы и правила сосуществования людей в обществе, наносящее вред объектам нападения (одушевленным и неодушевленным), причиняющее ущерб или вызывающее у них психологический дискомфорт: отрицательные переживания, состояния напряженности, страха, подавленности и др.» [Ениколопов С. Н. Понятие агрессии в современной психологии //Прикладная психология. – 2001. – №. 1. – С. 60-72.]. В данном определении

необходимо обратить особое внимание на адресность и целенаправленность деструктивного действия на одушевленный или неодушевленный объект. Этот ключевой признак выражает то, что без объекта нападения или жертвы агрессивное действие перестает быть агрессивным. Поэтому, поведение оценивается как агрессивное только в том случае, если в действии объект нападения или явно присутствует, или подразумевается, т.е. из контекста ситуации понятно, что жертва агрессивного акта есть, но не присутствует в кадре. Например, человек произносит адресные оскорбления другого человека (стример начинает ругаться с чатом или с кем-то в комнате, кто не попадает в кадр), находящегося за кадром, либо агрессор начинает бить человека, не попавшего в кадр.

В общем случае, модель, описывающая поведение в рамках взаимодействия «Агрессор-Жертва» приведена в разделе, посвященном разработке онтологическим моделям. Коротко, это ситуации, когда: агрессор нападает, жертва ведет себя пассивно; агрессор нападает, жертва отвечает агрессией; невозможно четко выделить агрессора и жертву, т.к. они нападают друг на друга одновременно.

В этом же разделе, а также в разделе, посвященном разработке классификаций, описаны все рассматриваемые виды агрессии и их направленности. Коротко, в данном исследовании рассматриваются физический и вербальный виды агрессии, а также следующие направления агрессии: на себя, на неодушевленные предметы, на другие одушевленные предметы. Все возможные сочетания этих видов и направлений составляют критерии оценки агрессивного поведения, которые можно выразить в виде следующей иерархической структуры:

1. Физическая агрессия:

1.1. На людей:

1.1.1. Удары руками, ногами и предметами

1.1.2. Кидание предметов

1.1.3. Толчки руками

1.1.4. Борцовские приемы:

1.1.4.1. Захваты

1.1.4.2. Подсечки

1.1.4.3. Удушающие приемы

1.1.4.4. Поваливание на пол

1.1.5. Таскание за волосы

1.1.6. Выкручивание носа

1.1.7. Удары головой

- 1.1.8. Обливание жидкостью
- 1.1.9. Прыжки на человека
- 1.1.10. Прочие
- 1.2. На себя:
 - 1.2.1. Удары руками и предметами
 - 1.2.2. Удары головой о предметы
 - 1.2.3. Порезы
 - 1.2.4. Удушения
 - 1.2.5. Стрельба в себя
 - 1.2.6. Прочие
- 1.3. На предметы:
 - 1.3.1. Удары руками, ногами, другими предметами
 - 1.3.2. Кидание предметов
 - 1.3.3. Укусы предметов
 - 1.3.4. Поломка
 - 1.3.5. Прыжки на предметы
 - 1.3.6. Прочие
- 2. Вербальная агрессия:
 - 2.1. Матерные или не матерные оскорбления
 - 2.2. Угрозы жизни, здоровью, благополучию
 - 2.3. Посылания
 - 2.4. Обвинения
 - 2.5. Приказы замолчать
 - 2.6. Пожелания смерти
 - 2.7. Поощрения насилия
 - 2.8. Оскорбления по гендерному, возрастному, расовому признакам
 - 2.9. Неприличные жесты
 - 2.10. Прочие

Примечание: в каждую категорию включен пункт «прочие». К нему относятся все те проявления, которые не попали в список, но которые однозначно можно трактовать как проявление агрессии.

Кроме этого, также необходимо выделить сопутствующее поведение, которое агрессией не является, но очень часто сопровождает агрессивные действия. Это эмоциональные и поведенческие проявления, к которым относятся:

1. Невербальное речевое поведение:
 - 1.1. повышенный голос,
 - 1.2. крики,
 - 1.3. вопли,
 - 1.4. плач,
 - 1.5. рычание
2. Двигательная активность тела, которая не наносит никому повреждений
 - 2.1. Сильная жестикуляция
 - 2.2. Резкие движения головой
 - 2.3. Резкие движения телом
3. Мимика или выраженные лицевые гримасы

Важно отметить, что конкретные агрессивные действия, входящие в состав приведенной иерархии, были выделены на основании статистического анализа результатов предварительного поискового наблюдения, описанного в [Uzdiaev M., Vatamaniuk I. Investigation of Manifestations of Aggressive Behavior by Users of Sociocyberphysical Systems on Video //Proceedings of the Computational Methods in Systems and Software. – Springer, Cham, 2021. – С. 593-604.]. Необходимо также отметить, что все описанные выше действия не исключают друг друга и могут выполняться одновременно в определенном временном промежутке.

Критерии и протокол оценки агрессивного поведения

В итоге, с учетом принятого нами разделения на физическую и вербальную агрессию, а также с учетом модели взаимодействия «АГРЕССОР-ЖЕРТВА», были выбраны следующие категории или критерии оценки, основанные на определенных выше видах агрессии и соответствующих им агрессивных действий:

1. АГРЕССОР: физическая агрессия на других людей
 - Любое из проявлений, описанных выше в перечне действий класса Физическая агрессия на других людей
2. АГРЕССОР: физическая агрессия на себя
 - Любое из проявлений, описанных выше в перечне действий класса Физическая агрессия на себя
3. АГРЕССОР: физическая агрессия на предметы
 - Любое из проявлений, описанных выше в перечне действий класса Физическая агрессия на предметы

4. АГРЕССОР: вербальная агрессия на других людей
 - Любое из проявлений, описанных выше в перечне действий класса Вербальная агрессия, которые человек адресует другому человеку
5. АГРЕССОР: вербальная агрессия на себя
 - Любое из проявлений, описанных выше в перечне действий класса Вербальная агрессия, которые человек адресует себе
6. АГРЕССОР: сопутствующее поведение, не являющееся агрессией
 - Сюда следует отнести все дополнительные эмоциональные и поведенческие проявления, описанные выше в перечне сопутствующих эмоциональных и поведенческих проявлений.
7. ЖЕРТВА: физическая агрессия на других людей (жертва дает сдачи агрессору или агрессорам)
 - Любое из проявлений, описанных выше в перечне действий класса Физическая агрессия на других людей
 - Следует размечать по этой категории только если в промежутке жертва дает сдачи агрессору, или они одновременно дерутся и непонятно, кто агрессор, а кто жертва
8. ЖЕРТВА: вербальная агрессия на других людей (жертва отвечает оскорблениями на оскорбления)
 - Любое из проявлений, описанных выше в перечне действий класса Вербальная агрессия, которые человек адресует другому человеку
9. ЖЕРТВА: сопутствующее поведение, не являющееся агрессией
 - Сюда следует отнести все дополнительные эмоциональные и поведенческие проявления, описанные выше в перечне сопутствующих эмоциональных и поведенческих проявлений.

Здесь особо пристальное внимание необходимо уделить тому факту, что обозначенные выше критерии не исключают друг, т.е. агрессивные действия различных типов могут выполняться одновременно на одном и том же временном интервале. Из этого следует, что каждый временной интервал, на который видео может быть разбито в ходе разметки, должен оцениваться независимо на соответствие или несоответствие (бинарная оценка) каждому из критериев.

Описанные выше критерии можно свести в таблицу, которую, в свою очередь, можно использовать в качестве протокола оценки фрагментов видео, на котором присутствует агрессивное поведение (таблица 1).

Таблица 1. Протокол оценки агрессивного поведения

Критерий оценки	Номер фрагмента видео			
	1	2	...	N
АГРЕССОР: физическая агрессия на других людей				
АГРЕССОР: физическая агрессия на себя				
АГРЕССОР: физическая агрессия на предметы				
АГРЕССОР: сопутствующее поведение, не являющееся агрессией				
ЖЕРТВА: физическая агрессия на других людей				
ЖЕРТВА: вербальная агрессия на других людей				
ЖЕРТВА: сопутствующее поведение, не являющееся агрессией				

В данной таблице, помимо критериев, также приведены временные интервалы, на которые разбивается видео. В этом случае выполняется оценка поведения по каждому из критериев на указанном временном интервале. При этом разметка может выполняться, во-первых, по принципу «хотя бы один раз» (сколько бы действий не встретилось, интервалу присваивается единственная оценка), а во-вторых, для каждого интервала может указываться количество искомых действий, которые встретились на указанном интервале. В общем случае, такие интервалы могут иметь произвольную длину и временную локализацию, однако, с целью унификации оценки оцениваемое видео может быть также разбито на последовательные интервалы одинаковой продолжительности по времени. Такой подход является типичным для оценки поведения в биологических, психологических и социальных науках [PALMER C. FRANK J. FLOYD, DONALD H. BAUCOM, JACOB J. GODFREY, and //Comprehensive Clinical Psychology. Volume 3. – 2000. – С. 1.].

Помимо этого, необходимо обозначить другие важные замечания:

1. Для того, чтобы успешно осуществить разметку большого количества видеоданных, необходимо составить протокол наблюдения, содержащий адекватное количество критериев оценки. Таким образом были сформированы приведенные выше девять критериев оценки поведения. По этой же причине, сопутствующие эмоциональные и поведенческие проявления укрупнены в одну категорию оценивания;
2. В случае разметки поведения жертвы не рассматриваются следующие ситуации: агрессия жертвы по отношению к себе и агрессия жертвы по отношению к неодушевленным предметам, т.к., данные ситуации избыточны и полностью покрываются критериями, оценивающими поведение агрессора;
3. В настоящем исследовании используется разбиение на одинаковые временные интервалы и оценка каждого интервала согласно обозначенным критериям по принципу «хотя бы один раз»;

4. Ввиду того, что обозначенные критерии не исключают друг друга, а соответствующие им типы поведения могут часто друг друга сопровождать (например, проявления физической агрессии сопровождается вербальной), каждый фрагмент необходимо оценивать по каждому отдельному критерию независимо от других критериев. Другими словами, в каждом временном интервале необходимо ставить отметку о наличии или отсутствии каждого вида поведения, входящего в критериальную сетку.

На основании разработанной методики оценки агрессивного поведения пользователей социкиберфизического пространства, представленного на видео, можно собрать и разметить выборку данных, для обучения на ней моделей машинного обучения, выполняющих распознавание агрессивных действий.

Сбор набора многомодальных данных, содержащих проявления агрессии в различных предметных ситуациях

Для сбора репрезентативной обучающей выборки необходимо применить описанную выше методику для оценки конкретных видео в открытом доступе в сети Интернет, содержащих агрессивное поведение пользователей социкиберфизического пространства. Процедура сбора и разметки набора данных состоит из следующих шагов:

1. Поиск и предварительный отбор видео соответствующей тематики в открытом доступе в сети Интернет на видеохостинговых платформах, таких как YouTube [<https://www.youtube.com/>]. В данном случае нас интересовало поведение пользователей т.н. «треш-стримов». Для отбора релевантных видео были использованы следующие поисковые запросы: «драка на стриме», «жесть на стриме», «треш на стриме»;
2. Далее каждое из отобранных видео было разбито на равные временные интервал по 10 с каждый. Интервал времени был выбран исходя из предполагаемого времени агрессивного действия, возможностей аннотаторов воспринимать оцениваемую информацию, а также исходя из общего предполагаемого времени аннотации всех видео;
3. Затем каждый из таких интервалов каждого из видео оценивается группой аннотаторов при помощи описанных в предыдущем разделе критериев, которые сведены в таблице 1. При этом, оценка производится по каждому из приведенных критериев по бинарному принципу наличия или отсутствия интересующего нас поведения. Также особо следует подчеркнуть, что разметка производится по

принципу «хотя бы один раз». Это означает, что размечаемый интервал помечается как содержащий искомое поведение, если это поведение встретилось хотя бы один раз. Даже в случае, если на размечаемом интервале встретилось более одного поведенческого проявления, для этого интервала выполняется единственная бинарная оценка;

4. Выполняется оценка согласованности разметки набора данных и окончательная разметка интервалов, получивших наиболее согласованные оценки.

Программный инструментарий разметки

Сама разметка и оценка выполнялись посредством программного инструментария ELAN [Sloetjes H., Wittenburg P. Annotation by category-ELAN and ISO DCR //6th international Conference on Language Resources and Evaluation (LREC 2008). – 2008.] версии 6.3. Это свободно распространяемое программное обеспечение (ПО), специально разработанное для выполнения аннотации поведения вербального и невербального поведения людей и животных, которое представлено на видео. Данное ПО имеет средства для просмотра видео, прослушивания аудио, разделение их на временные интервалы и оценка поведения в этих интервалах согласно необходимым для того или иного исследования критериям и системам критериев оценки. При этом, такие критерии оценки в ELAN можно гибко настраивать, что позволяет с легкостью перенести туда протокол из таблицы 1. Интерфейс ELAN приведен на рисунке 9.

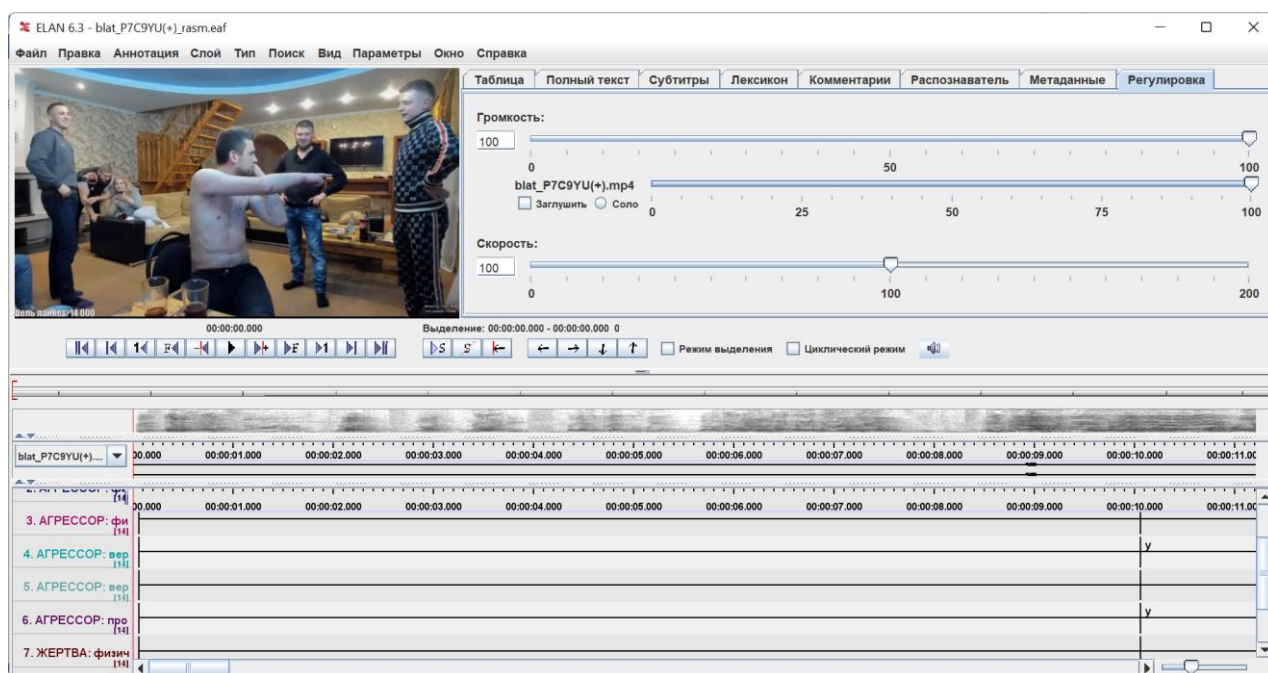


Рисунок 9. Интерфейс используемого для аннотации ПО ELAN

Для выполнения аннотации и сохранения ее результатов ELAN использует специальные файлы проектов, которые имеют синтаксис языка XML. В файлах-проектах ELAN помимо результатов аннотации также хранится система критериев оценки и информация о разбиении видео на временные интервалы. Кроме того, структурная организация языка XML файлов-проектов ELAN позволяет с легкостью извлекать из них результаты аннотации для каждого временного интервала.

Результаты отбора и аннотации видео

В результате поисковых запросов в видеохостинге YouTube было получено 138 видео, на которых потенциально содержится агрессивное поведение пользователей социкиберфизического пространства. Общая продолжительность видео составила 785 минут или 47100 десятисекундных интервалов, на которые эти видео были разделены для выполнения разметки.

Далее эти видео оценивались пятью аннотаторами возрастом 23-31 лет, трое мужчин, две женщины. Каждый аннотатор оценивал каждый десятисекундный интервал каждого видео согласно критериям, взятым из таблицы 1.

Для того, чтобы оценить надежность методики или воспроизводимость ее результатов в исследованиях обычно используется вычисление согласованности мнений экспертов, участвующих в оценке эмпирического материала. При этом, отмечается, что высокие показатели согласованности свидетельствуют о высокой надежности методики [Bakeman R., Quera V. Sequential analysis and observational methods for the behavioral sciences. – Cambridge University Press, 2011.][Hartmann D. P., Wood D. D. Observational methods //International handbook of behavior modification and therapy. – Springer, Boston, MA, 1990. – С. 107-138.]. Для оценки согласованности разметки в настоящем исследовании использовался коэффициент α Криппендорфа, который рассчитывается по формуле для каждой категории оценивания:

$$\alpha = 1 - \frac{D_o}{D_e}, \quad (2)$$

где D_o – расхождение в оценках, полученное в результате аннотации, D_e – оценка случайного расхождения в оценках. Выбор этого коэффициента обусловлен тем, что он включает в себя коррекцию случайного расхождения в оценках. Результаты вычисленных коэффициентов сведены в таблицу 2.

Таблица 2. Оценки согласованности разметки выборки посредством коэффициента α Криппендорфа

Категория	Значение коэффициента
АГРЕССОР: физическая агрессия на других людей	0.801
АГРЕССОР: физическая агрессия на себя	0.427
АГРЕССОР: физическая агрессия на предметы	0.642
АГРЕССОР: вербальная агрессия на других людей	0.834
АГРЕССОР: вербальная агрессия на себя	0.434
АГРЕССОР: сопутствующее поведение, не являющееся агрессией	0.622
ЖЕРТВА: физическая агрессия на других людей	0.792
ЖЕРТВА: вербальная агрессия на других людей	0.814
ЖЕРТВА: сопутствующее поведение, не являющееся агрессией	0.597

Как видно из таблицы, в большинстве случаев были получены высокие результаты оценок согласованности разметки пятью аннотаторами. Категории, для которых были получены низкие оценки, имеют малое количество экземпляров и незначительно влияют на результаты оценки. Полученные результаты коэффициентов согласованности экспертных оценок позволяют сделать вывод, о высокой надежности методики оценки, как уже было упомянуто выше для методик наблюдения за поведением. Для окончательного отбора экземпляров видео в обучающую выборку был использован метод голосования большинством: в выборку попали только те десятисекундные фрагменты, которые получили как минимум три одинаковые оценки аннотаторов из пяти.

В результате разметки была получена выборка, имеющая следующие параметры:

- Общее количество десятисекундных фрагментов с учетом всех рассматриваемых классов: 31788;
- Общее количество фрагментов, размеченных классом «АГРЕССОР: физическая агрессия на других людей»: 752;
- Общее количество фрагментов, размеченных классом «АГРЕССОР: физическая агрессия на себя»: 2;
- Общее количество фрагментов, размеченных классом «АГРЕССОР: физическая агрессия на предметы»: 30;
- Общее количество фрагментов, размеченных классом «АГРЕССОР: вербальная агрессия на других людей»: 986;
- Общее количество фрагментов, размеченных классом «АГРЕССОР: вербальная агрессия на себя»: 10;

- Общее количество фрагментов, размеченных классом «АГРЕССОР: сопутствующее поведение, не являющееся агрессией»: 95;
- Общее количество фрагментов, размеченных классом «ЖЕРТВА: физическая агрессия на других людей»: 375;
- Общее количество фрагментов, размеченных классом «ЖЕРТВА: вербальная агрессия на других людей»: 156;
- Общее количество фрагментов, размеченных классом «ЖЕРТВА: сопутствующее поведение, не являющееся агрессией»: 33.

Ввиду того, что у некоторых категорий в данной выборке слишком мало экземпляров, следующие категории исключаются из финальной выборки: «АГРЕССОР: физическая агрессия на себя», «АГРЕССОР: вербальная агрессия на себя», «АГРЕССОР: физическая агрессия на предметы», «АГРЕССОР: сопутствующее поведение, не являющееся агрессией», «ЖЕРТВА: сопутствующее поведение, не являющееся агрессией». (Для того, чтобы обеспечить приемлемый уровень распознавания, необходим дополнительный поиск и разметка видео, где содержатся конкретно эти категории). В результате в выборке осталось четыре класса:

- «АГРЕССОР: физическая агрессия на других людей»;
- «АГРЕССОР: вербальная агрессия на других людей»;
- «ЖЕРТВА: физическая агрессия на других людей»;
- «ЖЕРТВА: вербальная агрессия на других людей»;

При этом общее количество экземпляров в выборке составляет 2269 экземпляров, относящихся к обозначенным выше классам, а также 1000 экземпляров, на которых не присутствует агрессивное поведение.

Для выбора конкретной архитектуры нейронной сети, выполняющей многомодальное распознавание, необходимо выяснить, в каких соотношениях представлены модальности и их сочетания для различных видов агрессии в данной выборке: очевидно, что для редко встречающихся сочетания модальностей не будут вносить существенного вклада в результаты распознавания. Для этого необходимо вычислить, какие сочетания встречаются чаще всего. Это должно служить обоснованием выбора модальностей (вербальное и невербальное поведение), которые будут использованы для обучения и тестирования нейросетевой модели на данном наборе данных. Всего в выборке встретилось 65 сочетаний обозначенных классов, включая одиночные классы.

Приведем наиболее часто встречающиеся сочетания классов с указанием количества экземпляров, у которых эти сочетания встретились:

1. «АГРЕССОР: физическая агрессия на других людей» и «АГРЕССОР: вербальная агрессия на других людей»: 1395;
2. «АГРЕССОР: физическая агрессия на других людей» и «ЖЕРТВА: физическая агрессия на других людей»: 719;
3. «АГРЕССОР: физическая агрессия на других людей», «АГРЕССОР: вербальная агрессия на других людей» и «ЖЕРТВА: физическая агрессия на других людей»: 287;
4. «АГРЕССОР: вербальная агрессия на других людей» и «ЖЕРТВА: вербальная агрессия на других людей»: 176;
5. «АГРЕССОР: вербальная агрессия на других людей» и «АГРЕССОР: прочие эмоциональные и поведенческие проявления»: 147;
6. «АГРЕССОР: физическая агрессия на предметы» и «АГРЕССОР: вербальная агрессия на других людей»: 53
7. «АГРЕССОР: физическая агрессия на других людей», «АГРЕССОР: вербальная агрессия на других людей», «ЖЕРТВА: физическая агрессия на других людей» и «ЖЕРТВА: вербальная агрессия на других людей»: 42;
8. «АГРЕССОР: физическая агрессия на других людей», «АГРЕССОР: вербальная агрессия на других людей» и «АГРЕССОР: прочие эмоциональные и поведенческие проявления»: 39;
9. «АГРЕССОР: физическая агрессия на других людей», «АГРЕССОР: вербальная агрессия на других людей» и «ЖЕРТВА: вербальная агрессия на других людей»: 35;
10. «АГРЕССОР: физическая агрессия на других людей» и «ЖЕРТВА: вербальная агрессия на других людей»: 34.

Количество остальных сочетаний не превысило 30 экземпляров. Выявленное количество сочетаний показывает, что рассматриваемые в данной выборке категории часто пересекаются друг с другом в одних и тех же временных интервалах. В этом случае становится целесообразным обучение нейросетевой модели одновременно всем возможным задачам, которые соответствуют распознаванию всех включенных в выборку классов. При таком подходе обучения нейронных сетей можно добиться лучших показателей эффективности, т.к. здесь начинают влиять эффекты аугментации данных и многозадачного обучения, когда нейронная сеть подстраивается под различные аспекты одних и тех же данных, что, в свою очередь, расширяет набор возможностей распознавания.

Обучение и тестирование нейросетевых моделей распознавания агрессии на новом наборе данных

Решаемая задача распознавания агрессии целесообразно разбить на несколько подзадач, распознавания отдельных видов агрессии с точки зрения модели взаимодействия «Агрессор-жертва». Причина этого была уже упомянута в разделах выше: различные виды агрессии не исключают друг друга и могут проявляться одновременно на одном и том же временном промежутке. Эти условия определяют и структуру обучающей выборки данных, и структуру нейросетевой модели, которая должна решать задачу распознавания каждого вида агрессивного поведения в отдельности для каждого экземпляра видео.

Рассмотрим подробнее базовую нейросетевую модель, на основании которой будет строиться новая модель, выполняющая распознавание агрессивного поведения в рамках взаимодействия агрессора и жертвы. В качестве такой базовой модели была взята нейросетевая модель многомодального распознавания человеческой агрессии, подробно описанная в [Уздяев М. Ю. Нейросетевая модель многомодального распознавания человеческой агрессии //Вестник КРАУНЦ. Физико-математические науки. – 2020. – Т. 33. – №. 4.] [Uzdiaev M. Methods of multimodal data fusion and forming latent representation in the human aggression recognition task //2020 IEEE 10th International Conference on Intelligent Systems (IS). – IEEE, 2020. – С. 399-403.]. Данная модель отличается возможностью обработки произвольного количества модальностей, в которых может проявляться поведение человека (конкретно в данной модели рассматриваются такие модальности как вербальное речевое поведение, невербальное речевое поведение, двигательная активность тела, мимика лица). Другой отличительной особенностью является новый способ связывания векторов признаков отдельных модальностей для формирования единого признакового пространства, которое инвариантно виду обрабатываемой модальности и зависит только от распознаваемого класса. Общая архитектура этой модели показана на рисунке 10.

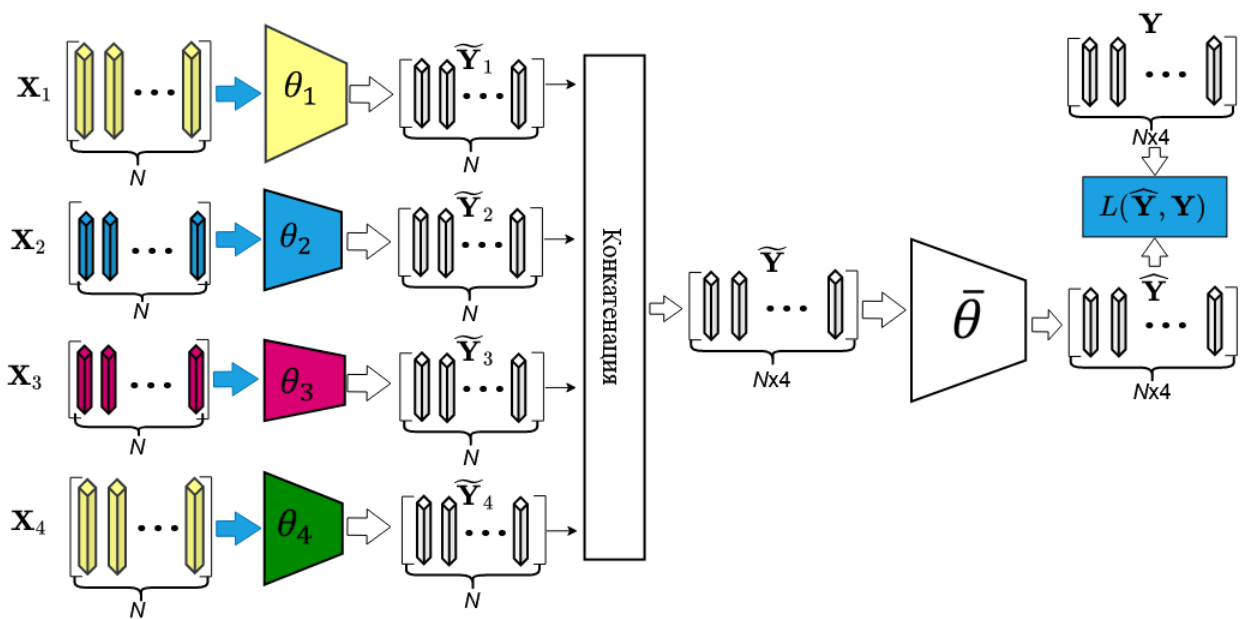


Рисунок 10. Архитектура нейросетевой модели многомодального распознавания агрессии

Рассмотрим архитектуру этой модели подробнее. Обработка данных различных модальностей моделью выполняется в два этапа: сначала по отдельности обрабатываются данные $\mathbf{X}_j$ разных модальностей соответствующими нейронными сетями с параметрами θ_j , которые формируют на своем выходе промежуточное признаковое представление $\tilde{\mathbf{Y}}_j$, где j – порядковый номер модальности. Промежуточное представление $\tilde{\mathbf{Y}}_j$ необходимо для того, чтобы обеспечить инвариантность признакового представления относительно вида модальности и сделать его зависимым только от распознаваемого класса. Полученное промежуточное представление $\tilde{\mathbf{Y}}_j$ подается на вход нейронной сети с параметрами $\bar{\theta}$, которая выполняет итоговую классификацию каждого экземпляра $\tilde{\mathbf{Y}}_j$ соответствующей модальности, формируя на выходе вектор вероятностей прогнозов распознанных классов $\hat{\mathbf{Y}}_i$. При этом инвариантность результатов распознавания относительно вида модальности обеспечивается в ходе процесса обучения, который заключается в том, что выходной нейросетевой классификатор настраивает свои параметры $\bar{\theta}$ на основании обработки всех экземпляров данных всех модальностей при помощи метода обратного распространения ошибки. В свою очередь, параметры нейронных сетей, выполняющих обработку отдельных модальностей θ_j также обновляются на основании ошибки, получаемой на выходе классификатора с параметрами $\bar{\theta}$ и распространяемой на эти нейронные сети с параметрами θ_j .

Стоит также особое внимание уделить способу обучения этой модели. Обучение описываемой модели выполняется при помощи пакетного стохастического градиентного

спуска с пакетами малого размера (Mini-Batch Stochastic Gradient Descend). Это означает, что нейронной сетью обрабатывается не отдельные экземпляры данных, а т.н. пакеты, состоящие из нескольких экземпляров, для каждого экземпляра в пакете формируется отдельный вывод, а значения градиентов функции потерь суммируются по обрабатываемому пакету данных. Именно применение такого типа организации обработки данных открывает возможности для формирования инвариантных относительно модальности промежуточных представлений $\tilde{\mathbf{Y}}$.

Пакетный градиентный спуск подразумевает одновременную обработку нейросетью не отдельных экземпляров данных, а т.н. пакетов данных, состоящих из множества отдельных экземпляров данных, случайно отобранных из обучающей выборки. При этом, выходная ошибка нейронной сети, получаемая на выходе, усредняется для всего обрабатываемого пакета. Таким образом, при обработке данных таким образом происходит учет и сглаживание их неоднородности, что в итоге приводит к повышению устойчивости обучения на сложных выборках большого объема.

Обработка многомодальной информации выполняется следующим образом: на вход каждой нейронной сети поступает пакет данных соответствующей модальности $\mathbf{X}_j$, который формируется путем конкатенации отдельных экземпляров данных этой модальности $\mathbf{X}_j = [\mathbf{x}_{j1}, \mathbf{x}_{j2}, \dots, \mathbf{x}_{jN}]$, где $\mathbf{X}_j \in \mathbb{R}^{d_j \times N}$, j – индекс модальности, N – размер пакета, d_j – размерность данных j -й модальности. На выходе каждой нейронной сети θ_j формируется пакет соответствующих промежуточных представлений $\tilde{\mathbf{Y}}_j = [\tilde{y}_{j1}, \tilde{y}_{j2}, \dots, \tilde{y}_{jN}]$. Далее формируется общий пакет промежуточных представлений с помощью конкатенации пакетов промежуточных представлений отдельных модальностей $\tilde{\mathbf{Y}} = [\tilde{\mathbf{Y}}_1, \tilde{\mathbf{Y}}_2, \dots, \tilde{\mathbf{Y}}_M]$, где M – количество обрабатываемых модальностей. Далее $\tilde{\mathbf{Y}}$ подается на вход нейросетевого классификатора с параметрами $\bar{\theta}$, который формирует на выходе пакет $\hat{\mathbf{Y}} = [\hat{\mathbf{Y}}_1, \hat{\mathbf{Y}}_2, \dots, \hat{\mathbf{Y}}_M]$, содержащий результаты распознавания для всех модальностей $\hat{\mathbf{Y}}_j = [\hat{y}_{j1}, \hat{y}_{j2}, \dots, \hat{y}_{jN}]$. Для $\hat{\mathbf{Y}}$ вычисляется функция потерь перекрестная энтропия (cross-entropy loss function) по всем экземплярам результатов распознавания для всех модальностей:

$$L(\hat{\mathbf{Y}}, \mathbf{Y}) = - \sum_{i=1}^M \sum_{j=1}^N \sum_{k=1}^C y_{ijk} \cdot \log(\hat{y}_{ijk}), \quad (3)$$

где M – количество модальностей, N – размер пакета, C – количество классов. Учет всех модальной в функции потерь, которая определяет величину коррекции параметров как выходного классификатора $\bar{\theta}$, так и нейронной сети каждой отдельной модальности как раз и обеспечивает формирование инвариантного относительно модальности промежуточного

признакового пространства $\tilde{\mathbf{Y}}$. Подробнее процесс обучения предлагаемой нейросетевой модели приведен в псевдокоде ниже.

Псевдокод алгоритма обучения:

Ввод:

K – количество пакетов, на которые разбивается обучающая выборка

M – количество модальностей

θ_j – веса нейронной сети, выполняющей обработку j -й модальности.

$\bar{\theta}$ – the parameters of the projection NN

```

1. while  $L(\mathbf{y}_{ij}, \hat{\mathbf{y}}_{ij}) \neq \min$  do:
2.   for  $i = 1 \dots K$ :
3.     for  $j = 1 \dots M$ :
4.       Получить пакет данных  $\mathbf{X}_j$ 
5.       Выполнить прямое распространение НС  $j$ -й модальности
            $\tilde{\mathbf{Y}}_j = f(\theta_j | \mathbf{X}_j)$ 
6.
7.     end
8.     Выполнить конкатенацию промежуточных представлений всех
           модальностей
            $\tilde{\mathbf{Y}} = [\tilde{\mathbf{Y}}_1, \tilde{\mathbf{Y}}_2, \dots, \tilde{\mathbf{Y}}_M]$ 
9.     Выполнить прямое распространение для выходного
           классификатора  $\hat{\mathbf{Y}} = f(\bar{\theta} | \tilde{\mathbf{Y}})$ 
10.    Вычислить ошибку  $L(\hat{\mathbf{Y}}, \mathbf{Y})$ 
11.    Обратное распространение ошибки и обновление весов
           выходного классификатора  $\bar{\theta} = \bar{\theta} - \eta \nabla_{\bar{\theta}} L(\hat{\mathbf{Y}}, \mathbf{Y})$ 
12.    for  $j = 1 \dots M$ :
13.      Выполнить обратное распространение ошибки  $L(\hat{\mathbf{Y}}, \mathbf{Y})$  и
           обновить веса НС  $j$ -й модальности  $\theta_j = \theta_j - \eta \nabla_{\theta_j} L(\hat{\mathbf{Y}}, \mathbf{Y})$ 
14.    end
15.  end
16. end

```

Для выполнения многомодального распознавания выполняется усреднение оценок прогнозируемых меток классов $\hat{\mathbf{y}}_i$ по всем представленным модальностям

$$\hat{\mathbf{y}} = \frac{1}{M} \sum_{i=1}^M \hat{\mathbf{y}}_i, \quad (4)$$

где M – количество модальностей.

В качестве нейронных сетей, выполняющих выделение признаков каждой отдельной модальности ($\theta_1 - \theta_4$) были использованы следующие архитектуры:

1. Нейронная сеть с параметрами θ_1 служит для выделения признаков невербального речевого поведения в аудиосигнале. В данной работе была использована нейронная сеть архитектуры VGG [Simonyan K., Zisserman A. Very deep convolutional networks for large-scale image recognition //arXiv preprint arXiv:1409.1556. – 2014.], обученная при помощи метода многозадачного обучения. Подробнее эта архитектура и метод обучения представлены в [Рябинов А. В., Уздяев М. Ю., Ватаманюк И. В. Применение многозадачного глубокого обучения в задаче распознавания эмоций в речи //Известия Юго-Западного государственного университета. – 2021. – Т. 25. – №. 1. – С. 82-109.].
2. Нейронная сеть с параметрами θ_2 служит для выделения признаков мимических изменений в видео. В данной работе был использован каскад сверточной нейронной сети архитектуры Sufflenet [Zhang X. et al. Shufflenet: An extremely efficient convolutional neural network for mobile devices //Proceedings of the IEEE conference on computer vision and pattern recognition. – 2018. – С. 6848-6856.], которая служит для выделения мимических признаков на отдельных кадрах, содержащих лица и трехмерных сверточных нейросетевых блоков для обработки пространственно-временной динамики мимических признаков. Архитектура этой модели приведена на рисунке 11.

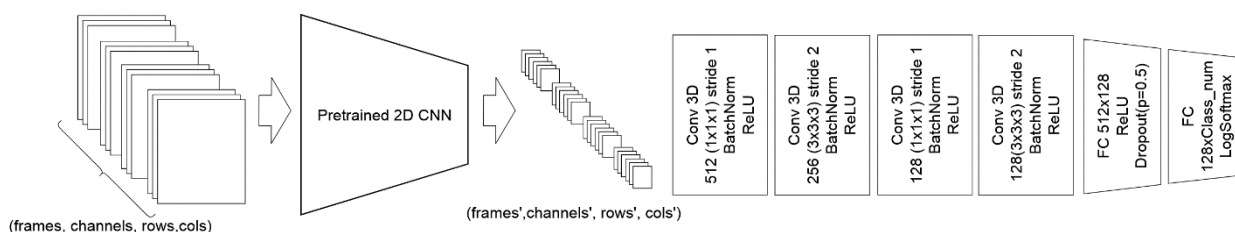


Рисунок 11. Архитектура нейронной сети, выполняющая выделение мимических признаков

3. Нейронная сеть с параметрами θ_3 служит для выделения признаков агрессии по вербальному речевому поведению. Или в текстовой расшифровке того, что говорит пользователь на видео. В данной работе была использована нейронная сеть архитектуры BERT [Devlin J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding //arXiv preprint arXiv:1810.04805. – 2018.], предварительно обученная на русскоязычных текстах и дообученная в рамках этой работы на собранной выборке.
4. Нейронная сеть с параметрами θ_4 служит для выделения признаков агрессии по двигательной активности тела, которая представлена на видео. В данной работе использовалась трехмерная сверточная нейронная сеть архитектуры I3D [Carreira J.,

Zisserman A. Quo vadis, action recognition? a new model and the kinetics dataset //proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. – 2017. – С. 6299-6308.]с дополнительными блоками, реализующими механизм нейросетевого внимания [Bahdanau D., Cho K., Bengio Y. Neural machine translation by jointly learning to align and translate //arXiv preprint arXiv:1409.0473. – 2014.]. Подробнее этот подход описан в работе [Uzdiaev M. Aggressive Action Recognition Using Pre-trained 3D CNNs with Attention //Proceedings of the Computational Methods in Systems and Software. – Springer, Cham, 2021. – С. 933-944.].

Отличительной особенностью данной модели в рамках ее обучения и тестирования на описанной в предыдущем разделе выборке данных является изменение архитектуры нейронной сети, выполняющей итоговую классификацию ($\bar{\theta}$). Ввиду того, что структура задачи, решаемой в данной работе, подразумевает распознавание каждого из видов агрессии по отдельности друг от друга, классификатор с параметрами $\bar{\theta}$ преобразовывается следующим образом: одна нейронная сеть с единственным нейроном, реализующим бинарную классификацию, разделяется на девять нейросетевых бинарных классификаторов. Данная архитектура изображена на рисунке 12.

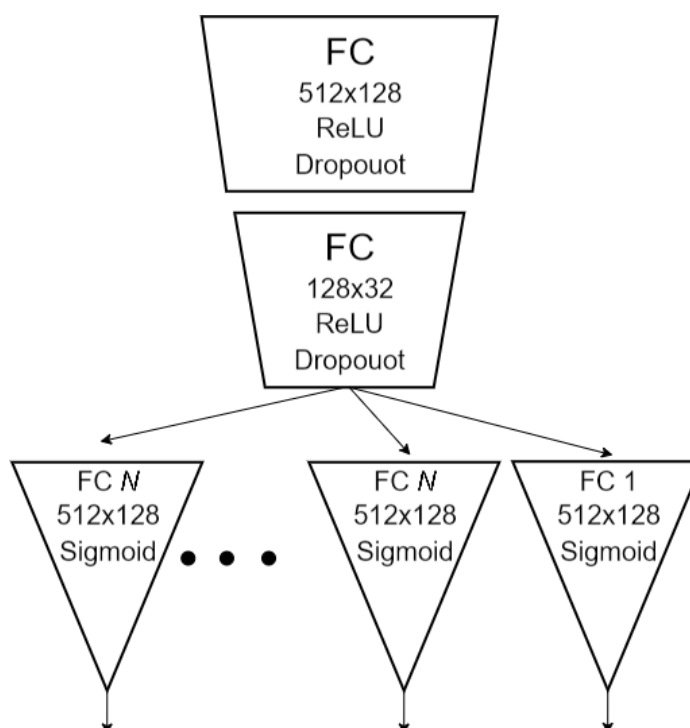


Рисунок 12. Архитектура выходного классификатора

Классификатор является трехслойной полносвязной нейронной сетью, у первых двух слоев которой активационная функция ReLU [Glorot X., Bordes A., Bengio Y. Deep sparse rectifier neural networks //Proceedings of the fourteenth international conference on

artificial intelligence and statistics. – JMLR Workshop and Conference Proceedings, 2011. – С. 315-323.], за которой следует механизм прореживания (dropout) [Srivastava N. et al. Dropout: a simple way to prevent neural networks from overfitting //The journal of machine learning research. – 2014. – Т. 15. – №. 1. – С. 1929-1958.] для обеспечения устойчивости процесса обучения. Выходные классификаторы для каждой задачи имеют сигмоидальную активационную функцию, которая и выполняет бинарную классификацию:

$$\sigma(z) = \frac{1}{1 + e^{-z}}, \quad (5)$$

где z – значение взвешенной суммы нейрона.

Организация эксперимента

Эксперимент организован следующим образом. Сначала выполняется обучение предлагаемой модели на обучающей выборке, состоящей из данных всех модальностей, представленных в размеченном в рамках исследования наборе данных: видео, содержащих мимические выражения (лицо), видео, содержащих двигательную активность тела (тело), невербального речевого поведения (звук), представленного в аудиосигнале, текста реплик пользователей (текст). При этом выполняется одновременное обучение всем задачам распознавания, представленным в выборке. Обучение выполнялось на обучающей выборке, полученной при помощи случайного отбора 80% экземпляров, входящих в изначальные 2269 экземпляров или 1815 экземпляров; для тестирования использовались остальные 20% изначальной выборки или 454 экземпляра. Упомянем задачи распознавания и количество экземпляров в обучающей и тестовой выборках:

1. Распознавание класса «АГРЕССОР: физическая агрессия на других людей» – 599 в обучающей и 153 в тестовой;
2. Распознавание класса «АГРЕССОР: вербальная агрессия на других людей» – 797 в обучающей и 189 в тестовой;
3. Распознавание класса «ЖЕРТВА: физическая агрессия на других людей» – 294 в обучающей и 81 в тестовой;
4. Распознавание класса «ЖЕРТВА: вербальная агрессия на других людей» – 125 в обучающей и 31 в тестовой;

Затем, на тестовой выборке выполняется тестирование обученной модели распознавать проявления агрессии на различных сочетаниях модальностей:

1. Лицо, тело, звук, текст
2. Лицо, тело, звук

3. Лицо, тело, текст
4. Лицо, звук, текст
5. Тело, звук, текст
6. Лицо, тело
7. Лицо, звук
8. Лицо, текст
9. Тело, звук
10. Тело, текст
11. Звук, текст

Таким образом, покрываются все возможные сочетания, которые могут быть в данном наборе модальностей. Обучение выполнялось в течение 200 эпох. Размер пакета в алгоритме пакетного градиентного спуска было выбрано равным 16. Для тестирования модели использовались настроенные параметры модели, полученные после прохождения 100 эпохи обучения. Это было обусловлено тем, что после 100 эпохи было достигнуто минимальное значение функции потерь на тестовой выборке. Обучение выполнялось с использованием библиотеки PyTorch [<https://pytorch.org/>].

Метрики

В данной работе будут использоваться следующие метрики.

1. Доля корректных прогнозов (accuracy).

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

2. Мера точности (precision).

$$precision = \frac{TP}{TP+FP} \quad (7)$$

3. Мера полноты (recall) или доли истинно положительных распознаваний (True Positive Rate – TPR)

$$recall = TPR = \frac{TP}{TP + FN} \quad (8)$$

4. F1-мера (F1-score) или среднее гармоническое precision и recall:

$$F1 - score = 2 \frac{precision \cdot recall}{precision + recall} = \frac{TP}{TP + \frac{TP + FN}{2}} \quad (9)$$

где TP – количество истинно положительных распознаваний, TN – количество истинной отрицательных распознаваний, FP – количество ложноположительных распознаваний, FN – количество ложноотрицательных распознаваний.

Результаты экспериментов и их обсуждение

Результаты распознавания класса «АГРЕССОР: физическая агрессия на других людей», достигнутые нейросетевой моделью для различных конфигураций модальностей приведены в таблице 3.

Таблица 3. Результаты экспериментального исследования распознавания класса «АГРЕССОР: физическая агрессия на других людей»

№	Конфигурация (сочетание модальностей)	Accuracy	Recall (TPR)	Precision	F1-score
1.	Лицо, тело, звук, текст	0.731	0.748	0.728	0.738
2.	Лицо, тело, звук	0.700	0.711	0.702	0.707
3.	Лицо, тело, текст	0.711	0.591	0.608	0.599
4.	Лицо, звук, текст	0.693	0.702	0.683	0.692
5.	Тело, звук, текст	0.718	0.704	0.738	0.721
6.	Лицо, тело	0.662	0.627	0.647	0.637
7.	Лицо, звук	0.599	0.486	0.497	0.491
8.	Лицо, текст	0.597	0.558	0.562	0.560
9.	Тело, звук	0.642	0.583	0.595	0.589
10.	Тело, текст	0.636	0.582	0.580	0.581
11.	Звук, текст	0.608	0.582	0.560	0.571
12.	Лицо	0.555	0.489	0.458	0.473
13.	Тело	0.615	0.569	0.524	0.546
14.	Звук	0.554	0.489	0.487	0.488
15.	Текст	0.566	0.435	0.488	0.460

Тестирование нейросетевой модели показало, что наилучшие результаты по общей доле верно распознанных экземпляров (ассигасу) показало сочетание всех представленных в выборке модальностей. Результаты метрики recall, отражающей способности алгоритма определять экземпляры искомого класса, также достигают лучших значений при учете всех модальностей. То же самое можно сказать про метрику f1-score. Метрика precision, отражающая способности алгоритма распознавать истинно положительные экземпляры среди всех положительных показала лучший результат у сочетания модальностей «Тело, звук, текст».

Эксперимент показал, что те сочетания (включая отдельные модальности), где есть модальность двигательной активности тела показывает лучшие результаты по сравнению с теми сочетаниями, где этой модальности нет. Это объясняется, в первую очередь, тем, что физическая агрессия проявляется в двигательной активности тела, которая отражена в видеопотоке.

Также стоит отметить, что при снижении количества обрабатываемых модальностей снижаются и результаты распознавания. Это экспериментально подтверждает более высокую эффективность распознавания при многомодальной обработке данных.

Результаты распознавания класса «АГРЕССОР: вербальная агрессия на других людей», достигнутые нейросетевой моделью для различных конфигураций модальностей приведены в таблице 4.

Таблица 4. Результаты экспериментального исследования распознавания класса «АГРЕССОР: вербальная агрессия на других людей»

№	Конфигурация (сочетание модальностей)	Accuracy	Recall (TPR)	Precision	F1-score
1.	Лицо, тело, звук, текст	0.752	0.757	0.760	0.759
2.	Лицо, тело, звук	0.723	0.694	0.720	0.706
3.	Лицо, тело, текст	0.735	0.718	0.730	0.724
4.	Лицо, звук, текст	0.735	0.729	0.736	0.732
5.	Тело, звук, текст	0.748	0.755	0.704	0.728
6.	Лицо, тело	0.679	0.634	0.664	0.648
7.	Лицо, звук	0.690	0.647	0.647	0.647
8.	Лицо, текст	0.719	0.712	0.703	0.707
9.	Тело, звук	0.688	0.686	0.650	0.668
10.	Тело, текст	0.708	0.704	0.660	0.681
11.	Звук, текст	0.720	0.719	0.719	0.719
12.	Лицо	0.556	0.581	0.538	0.559
13.	Тело	0.564	0.540	0.524	0.532
14.	Звук	0.625	0.624	0.633	0.629
15.	Текст	0.676	0.662	0.646	0.654

Тестирование нейросетевой модели показало, что также, как и в предыдущем случае, наилучшие результаты по общей доле верно распознанных экземпляров (ассигасу) показало сочетание всех представленных в выборке модальностей. Результаты метрики recall,

отражающей способности алгоритма определять экземпляры искомого класса, также достигают лучших значений при учете всех модальностей. То же самое можно сказать о результатах метрики precision, отражающей способности алгоритма распознавать истинно положительные экземпляры среди всех положительных. Метрика f1-score также показала лучший результат для сочетания всех модальностей.

Эксперимент выявил, что результаты распознавания выше там, где в сочетаниях есть текстовая модальность. Это объясняется тем, что вербальная агрессия проявляется в текстовой модальности, а остальные модальности являются сопутствующими ей, но ее не определяющими.

Эксперимент во второй раз показал, что при снижении количества обрабатываемых модальностей, снижаются также и результаты распознавания, что опять подтверждает обоснованность многомодальной обработки при распознавании.

Результаты распознавания класса «ЖЕРТВА: физическая агрессия на других людей», достигнутые нейросетевой моделью для различных конфигураций модальностей приведены в таблице 5.

Таблица 5. Результаты экспериментального исследования распознавания класса «ЖЕРТВА: физическая агрессия на других людей»

№	Конфигурация (сочетание модальностей)	Accuracy	Recall (TPR)	Precision	F1-score
1.	Лицо, тело, звук, текст	0.630	0.574	0.510	0.540
2.	Лицо, тело, звук	0.635	0.555	0.544	0.549
3.	Лицо, тело, текст	0.615	0.512	0.519	0.516
4.	Лицо, звук, текст	0.581	0.502	0.458	0.479
5.	Тело, звук, текст	0.625	0.532	0.568	0.550
6.	Лицо, тело	0.591	0.494	0.508	0.501
7.	Лицо, звук	0.525	0.344	0.346	0.345
8.	Лицо, текст	0.494	0.291	0.286	0.288
9.	Тело, звук	0.585	0.405	0.449	0.426
10	Тело, текст	0.599	0.557	0.515	0.535
11.	Звук, текст	0.544	0.325	0.346	0.335

12.	Лицо	0.455	0.252	0.266	0.259
13.	Тело	0.555	0.443	0.434	0.438
14.	Звук	0.484	0.281	0.258	0.269
15.	Текст	0.444	0.244	0.245	0.244

Тестирование нейросетевой модели показало, что наилучшие результаты по общей доле верно распознанных экземпляров (ассигасу) показало сочетание модальностей «лицо, тело, звук». Результаты метрики recall, отражающей способности алгоритма определять экземпляры искомого класса, достигают лучших значений при учете всех модальностей. Метрика precision, отражающая способности алгоритма распознавать истинно положительные экземпляры среди всех положительных показала лучший результат у сочетания модальностей «тело, звук текст». То же самое можно сказать про метрику f1-score.

Также, как и в задаче распознавания класса «АГРЕССОР: физическая агрессия на других людей», данный эксперимент показал, что те сочетания модальностей, где есть модальность двигательной активности тела показывает лучшие результаты по сравнению с теми сочетаниями, где этой модальности нет, что объясняется, выраженностью физической агрессии в двигательной активности тела, которая отражена в видеопотоке.

Также стоит отметить, что при снижении количества обрабатываемых модальностей снижаются и результаты распознавания. Это опять подтверждает более высокую эффективность распознавания при многомодальной обработке данных.

Результаты распознавания класса «ЖЕРТВА: вербальная агрессия на других людей», достигнутые нейросетевой моделью для различных конфигураций модальностей приведены в таблице 6.

Таблица 6. Результаты экспериментального исследования распознавания класса «ЖЕРТВА: вербальная агрессия на других людей»

№	Конфигурация (сочетание модальностей)	Accuracy	Recall (TPR)	Precision	F1-score
1.	Лицо, тело, звук, текст	0.545	0.412	0.354	0.381
2.	Лицо, тело, звук	0.560	0.375	0.441	0.405
3.	Лицо, тело, текст	0.585	0.406	0.400	0.403

4.	Лицо, звук, текст	0.595	0.465	0.434	0.449
5.	Тело, звук, текст	0.580	0.500	0.417	0.455
6.	Лицо, тело	0.490	0.288	0.257	0.271
7.	Лицо, звук	0.495	0.258	0.225	0.241
8.	Лицо, текст	0.550	0.469	0.350	0.458
9.	Тело, звук	0.495	0.366	0.317	0.340
10	Тело, текст	0.550	0.386	0.451	0.416
11.	Звук, текст	0.565	0.444	0.447	0.392
12.	Лицо	0.485	0.156	0.169	0.163
13.	Тело	0.475	0.316	0.214	0.255
14.	Звук	0.525	0.449	0.352	0.395
15.	Текст	0.555	0.456	0.373	0.411

Тестирование нейросетевой модели показало, что наилучшие результаты по общей доле верно распознанных экземпляров (ассигасу) показало в сочетании «лицо, звук, текст». Результаты метрики recall, отражающей способности алгоритма определять экземпляры искомого класса, достигают лучших значений в сочетании «тело, звук, текст». Метрика precision, отражающая способности алгоритма распознавать истинно положительные экземпляры среди всех положительных, показала лучший результат у сочетания модальностей «звук текст». Метрика f1-score показала лучший результат у сочетания «Тело, звук, текст». При этом, отмечается не высокие показатели метрик precision, recall, f1-score, что может быть объяснено малым размером выборки.

Эксперимент выявил, что результаты распознавания выше там, где в сочетаниях есть текстовая модальность. Это объясняется тем, что вербальная агрессия проявляется в текстовой модальности, а остальные модальности являются сопутствующими ей, но не определяющими.

Отметим выявленную уже тенденцию: при снижении количества обрабатываемых модальностей снижаются и результаты распознавания, что в очередной раз доказывает более высокую эффективность предлагаемого подхода многомодальной обработки данных.

В целом, также следует отметить, то при распознавании класса «ЖЕРТВА: вербальная агрессия на других людей» были достигнуты самые низкие показатели. Это, в первую очередь, объясняется малой выборкой, которая была использована для обучения и тестирования алгоритмов.

Выводы по всему исследованию

В рамках всего исследования был разработан комплексный междисциплинарный подход к анализу и распознаванию агрессии русскоязычных пользователей информационного пространства. Междисциплинарность подхода заключается в использовании методов, методик и подходов различных научных дисциплин: психологии, математической статистике, теории машинного обучения и нейронных сетей. При этом знания различных научных дисциплин использовались в соответствии или системным представлением о решаемой задаче распознавания агрессивного поведения, что обеспечивает комплексность разработанного подхода. При этом системное представление, в соответствии с которым использовались подходы различных дисциплин, подчиняется следующему принципу, выражаемому в конкретных шагах:

1. Выполняется детальный анализ предметной области проявлений человеческой агрессии, представленных на видео, в информационном пространстве. На основе этого анализа, а также предварительных теоретических сведений и наиболее распространенных классификаций агрессии, взятых из психологии, выделяются наиболее значимые признаки агрессии. В рамках этого же шага, на основании предварительного поискового наблюдения за поведением, представленном на видео в открытом доступе, русскоязычных пользователей информационного пространства уточняется специфика агрессивного поведения, особенности взаимодействия участников агрессивных актов, а также определяются конкретные действия, которые ассоциируются с агрессией. В результате формируются классификации агрессивного поведения, отражающие наиболее значимые признаки с учетом специфики проявлений агрессии именно русскоязычными пользователями в информационном пространстве.
2. Выполняется формализация предметной области агрессии русскоязычных пользователей в информационном пространстве, отраженной в классификациях, посредством онтологических моделей, которые, в свою очередь, помимо классов, отражают также наиболее значимые отношения между этими классами. Также онтологические модели реализуют классификационные правила, по которым можно оценивать агрессивное поведение. В рамках этого же шага выполняется

разработка методики оценки агрессивного поведения. Методика разрабатывается на основании онтологических моделей, отражает наиболее значимые классы предметной области и отношения между классами. Обоснованность методики обеспечивается использованием общепсихологических представлений об агрессии, учете специфики предметной области и анализе предварительного поискового наблюдения, в результате которого были выделены конкретные действия, относящиеся к агрессивному поведению. Разработанная методика может быть использована для сбора и разметки выборки данных, которая, в свою очередь, может быть использована для обучения моделей машинного обучения, выполняющих распознавание агрессии.

3. Выполняются сбор и разметка видеоданных, находящихся в открытом доступе в сети Интернет и содержащих агрессивное поведение русскоязычных пользователей. При этом, в рамках данного исследования для разметки используется обоснованная методика оценки, которая упомянута выше, и в ходе разметки выборки выполняется апробация методики и проверка ее надежности посредством вычисления согласованности оценок экспертов. Размеченная выборка содержит различные виды агрессии, представленные в различных модальностях: двигательная активность тела на видео, вербальное речевое поведение в транскрибированном тексте, невербальное речевое поведение в аудио, мимические выражения на видео. Выборка может быть использована для обучения нейросетевых моделей, выполняющих одно- и многомодальное распознавание агрессивного поведения.
4. Выполняется разработка нейросетевых моделей одно- и многомодального анализа и распознавания агрессивного поведения. В рамках данной работы разработаны нейросетевые модели распознавания на основе анализа следующих модальностей: двигательная активность тела, вербальное речевое поведение, невербальное речевое поведение, мимические выражения. Разработанные модели также используются в рамках подхода переноса обучения для выполнения многомодального распознавания агрессии. Все разработанные модели обучаются и тестируются на обозначенной выше выборке данных, содержащей агрессивное поведение русскоязычных пользователей в информационном пространстве. В результате тестирования многомодальный подход показал лучшие результаты распознавания по сравнению с одномодальным (75.2% точности для многомодального распознавания вербальной агрессии против 67.6% точности для одномодального распознавания;

73.1% точности для многомодального распознавания физической агрессии против 61.5% точности для одномодального распознавания).

Комплексный подход, выраженный в описанных выше шагах, является целостным выражением предметной области агрессивного поведения, которая рассматривается с различных сторон посредством междисциплинарного анализа. Выведенные выше шаги являются базовыми или самыми общими составляющими разработанного комплексного междисциплинарного подхода. При этом, в ходе анализа конкретной предметной области выполняется конкретизация и уточнение каждого из шагов с учетом специфики предметной области (в нашем случае, проявлений агрессии русскоязычными пользователями в информационном пространстве). Однако, именно базовые составляющие этого подхода позволяют расширять и масштабировать его на решение смежных задач анализа и автоматического распознавания поведения человека.