DOI: 10.6060/ivecofin.2026671.774
УДК: 004.421

ПРОГРАММНАЯ СИСТЕМА ДЛЯ МУЛЬТИМОДАЛЬНОГО АНАЛИЗА ЭМОЦИОНАЛЬНОГО СОСТОЯНИЯ ЧЕЛОВЕКА НА ОСНОВЕ АНСАМБЛЕВЫХ МЕТОДОВ КЛАССИФИКАЦИИ

Д. В. Фазульянов, А.И. Гусева

Дмитрий Владимирович Фазульянов* (ORCID 0009-0004-0548-2982), Анна Ивановна Гусева (ORCID 0000-0002-7236-1257)
Национальный исследовательский ядерный университет «МИФИ», ш. Каширское, 31, Москва, 115409, Россия
Email: fazulianov.dmitrii@gmail.com*, aiguseva@mephi.ru

Данная статья посвящена разработке и исследованию программной системы для комплексного распознавания эмоциональных состояний человека на основе анализа видео, аудио и текстовых данных. Предложена архитектура, ориентированная на работу в условиях реального времени и использующая ансамблевые методы классификации для повышения надежности итоговых прогнозов. Архитектура системы построена на распределенном трехуровневом принципе, включающий клиентское приложение, высокопроизводительный API на базе FastAPI и изолированную сеть вычислительных узлов под управлением Celery и Redis. Методология исследования опирается на параллельное извлечение признаков: визуальных – через сверточную нейронную сеть ResNet-50, акустических характеристик с помощью энкодера LSTM и MFCC и семантических векторов с помощью Multilingual BERT. Ключевым элементом решения является стратегия слияния данных (Late Fusion) на основе ансамбля моделей, таких как: Stacking, Voting, Adaboost и CatBoost. Также внимание уделено механизму временной синхронизации на основе временных меток модели Faster-Whisper, что позволяет сопоставлять видеокадры, фрагменты речи и текстовую расшифровку, и минимизировать ошибки из-за задержек в обработке разных модальностей. В экспериментальной части проведено сравнение разработанного программного решения с классическими нейросетевыми архитектурами мультимодального сентимент-анализа Tensor Fusion Network (TFN) и Multi-attention Recurrent Network (MARN). Результаты апробации на реальных данных показывают, что предложенный подход на основе градиентного бустинга превосходит указанные аналоги по точности и устойчивости к внешним шумам. Реализованный подход обладает широким спектром применения, например, для мониторинга эмоционального климата в распределенных командах, что помогает своевременно выявлять риски выгорания и предупреждать конфликты. Также предложенная программная система позволяет автоматизировать аудит качества клиентского сервиса и оценку вовлеченности учащихся в рамках образовательных платформ.

Ключевые слова: сентимент-анализ, анализ эмоциональных состояний, Stacking, Adaboost, CatBoost, программная система, инженерное решение, ансамблевые методы классификации.

A SOFTWARE SYSTEM FOR MULTIMODAL ANALYSIS OF HUMAN EMOTIONAL STATE BASED ON ENSEMBLE CLASSIFICATION METHODS

D.V. Fazulianov, A.I. Guseva

Dmitry V.Fazulianov* (ORCID 0009-0004-0548-2982), Anna I. Guseva (ORCID 0000-0002-7236-1257)
National Research Nuclear University "MEPhI", Kashirskoe highway, 31, Moscow, 115409, Russia
Email: fazulianov.dmitrii@gmail.com*, aiguseva@mephi.ru

This article is devoted to the development and research of a software system for complex recognition of human emotional states based on the analysis of video, audio and text data. An architecture is proposed which focuses on real-time operation and uses ensemble classification methods to increase the reliability of final forecasts. The architecture of the system is based on a distributed three-level principle, including a client application, a high-performance API based on the Fast API, and an isolated network of computing nodes

running Celery and Redis. The research methodology is based on parallel extraction of features: visual – through the convolutional neural network ResNet-50, acoustic characteristics – using the LSTM and MFCC encoder, and semantic vectors – using Multilingual BERT. A key element of the solution is a Late Fusion strategy based on an ensemble of models such as Stacking, Voting, Adaboost, and CatBoost. Emphasis is also put on the time synchronization mechanism based on the timestamps of the Faster-Whisper model, which allows for the comparison of video frames, speech fragments and text transcription, minimizing errors due to delays in processing different modalities. Ehe experimental part describes a comparative study of the developed software solution with the classical neural network architectures of multimodal sentiment analysis (Tensor Fusion Network and Multi-attention Recurrent Network (MARN)). The results of tests on real data show that the proposed approach based on gradient boosting is superior to these analogues in terms of accuracy and resistance to external noise. The implemented approach has a wide range of applications, for example, for monitoring the emotional climate in teams, which helps to identify burnout risks in a timely manner and prevent conflicts. The proposed software system also makes it possible to automate the audit of customer service quality and the assessment of student engagement within educational platforms.

Keywords: sentiment analysis, emotional state analysis, Stacking, Adaboost, CatBoost, software system, engineering solution, ensemble classification methods.

.

Для цитирования:
Фазульянов Д.В., Гусева А. И. Программная система для мультимодального анализа эмоционального состояния человека на основе ансамблевых методов классификации. *Известия высших учебных заведений. Серия «Экономика, финансы и управление производством» [Ивэкофин]*. 2026. № 01(67). С. 192-202. DOI: 10.6060/ ivecofin.2026671.774

For citation:
Fazulianov D.V., Guseva A. I. A software system for multimodal analysis of human emotional state based on ensemble classification methods. *Ivecofin.* 2026. N 01(67). P. 192-202. DOI: 10.6060/ivecofin.2026671.774 (in Russian)

ВВЕДЕНИЕ

В условиях быстрого развития цифровой экономики и интеллектуальных сервисов задача автоматизированного распознавания эмоциональных состояний человека приобретает особую значимость. Анализ эмоций пользователей находит применение в различных технологических отраслях, позволяя улучшить взаимодействие с клиентами, оптимизировать процессы и повысить эффективность бизнеса. Основные направления использования включают финансы, здравоохранение, образование, маркетинг, HR-технологии, а также UX-дизайн и разработку программного обеспечения (ПО) [1-3]. Современные системы управления и рекомендательные сервисы стремятся не просто обрабатывать запросы пользователей, но и адаптироваться к их текущему психологическому настрою для повышения качества клиентского опыта и взаимодействия в цифровой среде.

На сегодняшний день в научном сообществе признано, что наиболее полная картина эмоционального фона может быть получена через мультимодальный сентимент-анализ – одновременную обработку видеоряда, аудиопотока и текста. Как отмечают авторы работы [4], мультимодальный анализ в российском исследовательском пространстве переходит от чисто теоретических концепций к поиску конкретных алгоритмов интеграции различных каналов восприятия.

Одной из важных задач, которая решается при обработке мультимодальных данных – объединение разнородных данных, полученных из разных источников. Основы интеграции разнородных данных были заложены в работах зарубежных исследователей, которые классифицировали методы слияния модальностей на ранние, поздние и гибридные [5, 6]. При раннем слиянии извлеченные признаки объединяются в единый высоко размерный вектор перед подачей на вход классификатору. Позднее слияние, напротив базируется на агрегации решений независимых моделей, обеспечивая высокую гибкость и устойчивость к зашумленным или неполным входным данным. Гибридные методы стремятся объединить эти подходы, используя многоуровневые архитектуры для выявления кросс-модальных зависимостей, чаще всего через attention-механизмы. Но на практике реализация таких систем сталкивается с рядом технологических вызовов. Одной из проблем является необходимость обеспечения высокой производительности при работе с «тяжелыми» нейросетевыми моделями в режиме реального времени.

Вопросы архитектурного построения подобных систем обсуждаются в современных отечественных работах. Авторы [7, 8] подчеркивают,

что при разработке веб-сервисов на основе нейросетей критически важным является разделение вычислительной нагрузки и использование специализированных библиотек для обработки аудио и видеосигналов.

Несмотря на наличие значительного количества наработок, в большинстве существующих решений модели часто отказываются чувствительны к шумам в данных, а также недостаточно проработан механизм синхронизации данных.

Целью данной работы является разработка и апробация программной системы, которая решает проблему надежности распознавания эмоций за счет использования ансамблевых методов и механизма временной синхронизации модальностей. Это позволяет системе эффективно работать даже в условиях неполных или зашумленных данных, что подтверждается сравнительным анализом с классическими нейросетевыми архитектурами.

МЕТОДОЛОГИЯ ИССЛЕДОВАНИЯ

Методология исследования опирается на мультимодальную архитектуру, объединяющую три независимых канала данных: видеоряд, аудиопоток и текстовую транскрипцию. Интеграция данных на уровне мета-признаков (позднее слияние, Late Fusion) позволяет решать проблему неполноты информации, типичную для реальных сценариев взаимодействия. Сочетание анализа мимики, позы тела, характеристик речи и семантического контекста обеспечивает более точную интерпретацию эмоционального фона по сравнению с мономодальными аналогами.

Процесс анализа строится на параллельной работе трех специализированных вычислительных модулей (визуальный, акустический и лингвистический) каждый из которых служит для извлечения специфических дескрипторов эмоционального состояния.

Визуальный модуль осуществляет покадровый анализ видеопотока с использованием сверточной нейронной сети ResNet-50 [9]. Основной задачей модуля является идентификация динамических изменений мимики и микровыражений лица, а также позы тела. Процесс извлечения признаков организован следующим образом.

1. Из видеопотока извлекается до 48 ключевых кадров с частотой 1 кадр/сек. Каждый кадр масштабируется и обрезается до размера 224x224 пикселя с последующей нормализацией по среднему значению и стандартному отклонению ImageNet.
2. Кадры подаются на вход ResNet-50, у которой удален финальный классификационный слой. Это позволяет получить дескриптор размерностью 2048 для каждого изображения.
3. Полученная последовательность векторов усредняется, формируя единый статический вектор визуальных признаков для всего сегмента видео.

Акустический модуль предоставляет два взаимозависимых метода анализа аудиодорожки, адаптируемых под вычислительные мощности и требования к точности. Процесс извлечения может быть описан следующим образом.

1. Алгоритм MFCC-240: реализует классический подход на основе мел-кепстральных коэффициентов. Из аудиосигнала извлекаются 40 коэффициентов MFCC, а также их первые и вторые производные. Для каждого из 120 параметров вычисляются среднее значение и стандартное отклонение по всей длине записи, что дает итоговый вектор из 240 признаков.
2. Энкодер LSTM: представляет собой обучаемую архитектуру для захвата временных зависимостей в речи. На вход подаются лог-мел-спектрограммы (64 канала). Архитектура включает двунаправленный двухслойный LSTM-блок с 128 скрытыми единицами. Финальный эмбеддинг формируется путем усреднения скрытых состояний по времени и последующей линейной проекции [10].

Лингвистический модуль выполняет семантический разбор текста, полученного в результате транскрибации речи моделью faster-whisper [11], обрабатывается моделью Multilingual BERT [12]. Это позволяет системе одинаково эффективно работать с контентом на разных языках. Процесс извлечения признаков работает следующим образом.

1. Текст разбивается на токены с ограничением длины до 256 единиц.
2. Вместо использования стандартного [CLS]-токена, в системе реализована стратегия усреднения состояний всех токенов с учетом маски внимания (Attention Mask). Это дает более устойчивое представление коротких фраз.
3. В случае отсутствия текста в сэмпле, функция возвращает нулевой вектор, который впоследствии игнорируется мета-классификатором за счет определенного флага.

Для обеспечения временной синхронизации разнородных сигналов (кадры видео, звуковые волны и текстовые токены) в системе реализован механизм синхронизации на основе временных меток, которые генерируются на этапе распознавания речи. Каждому текстовому токену сопоставляется интервал аудиоволны и набора видеокадров, зафиксированных в момент произнесения слова.

Для принятия итогового решения в системе применена стратегия поздней интеграции (Late Fusion). Данный подход предполагает, что каждый

информационный канал анализируется независимой моделью (SVM), после чего их выходные значения объединяются в единый вектор мета-признаков. Программная система поддерживает четыре алгоритмических подхода к интеграции данных, выбираемых пользователем через интерфейс:

- **Stacking** – логистическая регрессия, обучаемая на конкатенированном векторе вероятностей базовых моделей. Метод позволяет автоматически определить весовые коэффициенты доверия каждой модальности [13, 14];
- **Boosting** (CatBoost/AdaBoost) – использование градиентного и адаптивного бустинга для поиска нелинейных зависимостей между предсказаниями модальностей. Эти модели обучаются на расширенном наборе мета-признаков [15];
- **Soft Voting** – равновесное усреднение вероятностей, представляющее собой наиболее быстрый и интерпретируемый метод «коллективного мнения».

Использование вариативной мета-классификации позволяется адаптировать систему под конкретные эксплуатационные задачи: от высокоточного оффлайн-анализа (CatBoost / Stacking) до оперативного мониторинга эмоционального фона в реальном времени (Voting). Выбор мета-классификатора осуществляется на уровне конфигурации системы без изменения основной архитектуры конвейера обработки данных.

АРХИТЕКТУРА СИСТЕМЫ И ФУНКЦИОНАЛЬНЫЕ ВОЗМОЖНОСТИ

Проектирование программной системы проводилось на базе распределенной трехуровневой архитектуры. Такой подход позволяет изолировать пользовательский интерфейс от ресурсоемких вычислительных процессов, обеспечивая высокую масштабируемость и стабильную работу системы при одновременной обработке нескольких медиафайлов [16].

Архитектурное решение включает три основных уровня взаимодействия: клиентский уровень (Frontend), уровень оркестрации (Backend API) и уровень обработки. Архитектура программной системы представлена на рис. 1.

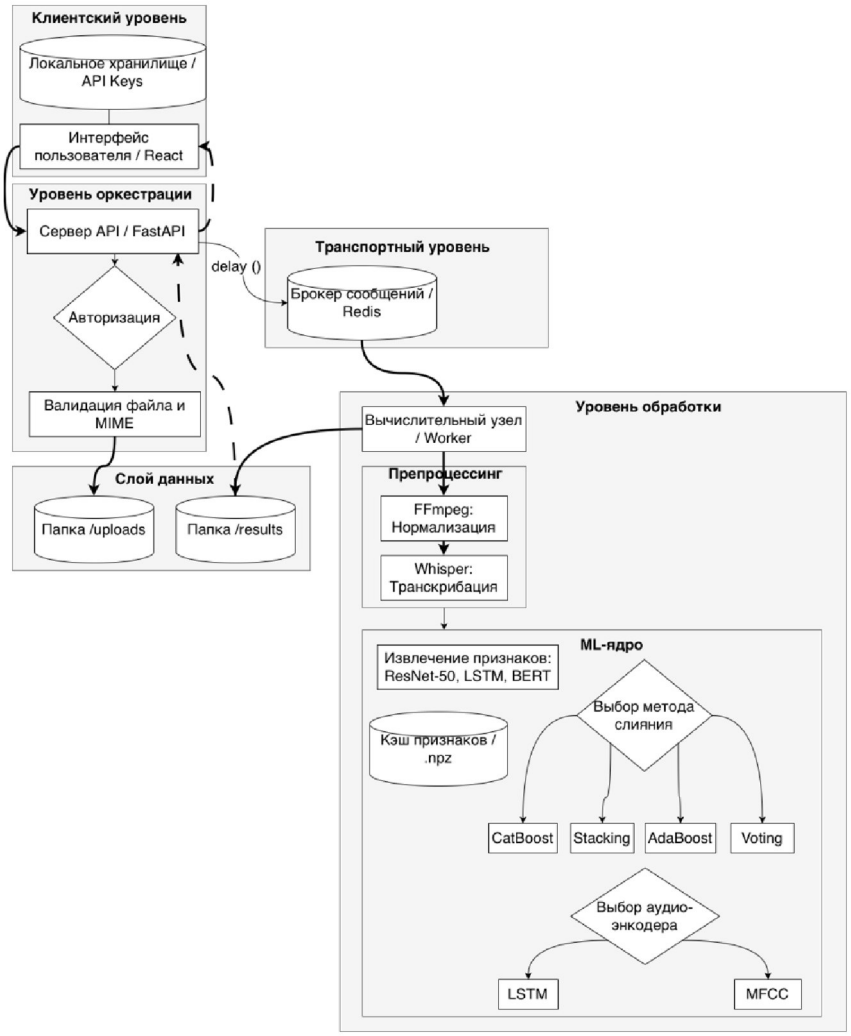

Рисунок 1. Архитектура программной системы
Figure 1. Software system architecture

Клиентский уровень (Frontend)

Интерфейс пользователя реализован в виде SPA-приложения (Single Page Application) с использованием компонентного подхода.

Основной задачей этого уровня является абстрагирование сложности внутренних аналитических процессов и предоставление интуитивно понятных инструментов управления конфигурацией. Помимо прямой настройки параметров анализа, на этом уровне реализованы механизмы первичной валидации данных.

В системе реализован механизм управления состоянием, позволяющий пользователю в реальном времени конфигурировать параметры задачи перед отправкой в очередь обработки. Функциональные возможности настройки включают следующие ключевые элементы:

- переключение между специализированными наборами весов моделей, такими как режим «Базовый» (обученном на выборках RAVDESS/EnterFace) и режим «Универсальный» (дополненный данными из набора CREMA-D) [17-19];
- возможность выбора аудио-энкодера между классическим анализом признаков (MFCC) и методами глубокого обучения на базе LSTM;
- функционал позволяет активировать параллельный запуск нескольких алгоритмов слияния (Stacking, Voting, CatBoost, AdaBoost) для одного файла, что обеспечивает возможность сравнительного анализа результатов в рамках одной вычислительной сессии;
- возможность выбора языка диктора через сочетание автоматического распознавания речи и ручного переключения на русский или английский язык.

Техническая реализация клиентской части включает ряд решений, повышающих стабильность и эргономику системы. Для сокращения времени подготовки к анализу при повторных сессиях используется локальное хранилище браузера (localStorage), в которому кэшируются параметры API-доступа (URL и ключ авторизации). Встроенная пре-валидация MIME-типов файлов, для видео и аудио на стороне клиента позволяет снизить нагрузку на сервер, отсекая некорректные запросы на раннем этапе. Интерфейс пользователя системы представлен на рис.2.

Уровень оркестрации (Backend API)

Центральным звеном системы выступает сервер приложений, функционирующий как интеллектуальный шлюз между клиентом и вычислительными мощностями. Реализация данного уровня на базе фреймворка FastAPI обусловлена его высокой производительностью, что позволяет эффективно управлять асинхронными потоками данных. На рис. 3 представлен жизненный цикл запроса от момента загрузки файла до получения результата. На диаграмме последовательности представлены основные участники процесса и этапы обработки пользовательского запроса.

1. Пользователь инициирует процесс, загружая медиафайл и выбирая параметры конфигурации.
2. React Client обеспечивает интерфейс взаимодействия, выполняя первичную отправку данных и реализует механизм периодического опроса для отслеживания состояния задачи.
3. FastAPI Server выполняет роль оркестратора, проверяет API-ключ и MIME-типы, регистрирует задачу в очереди и предоставляет данные о её текущем статусе и финальном результате.
4. Redis/Celery обеспечивают асинхронную передачу данных и управление очередью задач, позволяя серверу работать в неблокирующем режиме.
5. Worker (вычислительный узел), выполняющий непосредственную обработку ансамбля нейросетевых моделей и сохранение отчета в хранилище.

Основная роль Backend-составляющей заключается в комплексном управлении жизненным циклом аналитических задач. В системе реализованы следующие функции и инженерные решения.

1. Для обеспечения стабильности системы процесс передачи файлов происходит фрагментарно (порциями по 1 МБ). Это позволяет контролировать объем данных непосредственно в процессе загрузки и мгновенно прерывать соединение, если файл превышает допустимый лимит в 200 МБ. Такой подход предотвращает перегрузку памяти сервера и защищает его от внешних атак.
2. Система проверяет подлинность загружаемых данных, анализируя их внутреннюю структуру (цифровые подписи), а не только расширение файла. Это гарантирует, что на обработку нейросетям поступают только корректные медиафайлы.
3. Сервер функционирует в неблокирующем режиме, после получения запроса он регистрирует задачу в очереди и сразу возвращает пользователю уникальный идентификатор. Основная обработка данных нейросетями происходит в фоновом режиме, что позволяет пользователю отслеживать прогресс, не поддерживая постоянное активное соединение.
4. Архитектура поддерживает возможность одновременного тестирования различных комбинаций моделей на одном файле. Пользователь может задать параметры анализа, например, сравнить работу методов, и система автоматически распределит вычислительную нагрузку для получения сравнительных результатов в рамках одной операции.
5. Реализовано логическое и физическое разграничение статических ресурсов. Отдельная директория (uploads) используется для временного хранения оригинальных медиаданных, а другая директория (results) служит хранилищем для финальных структурированных отчетов в формате JSON.

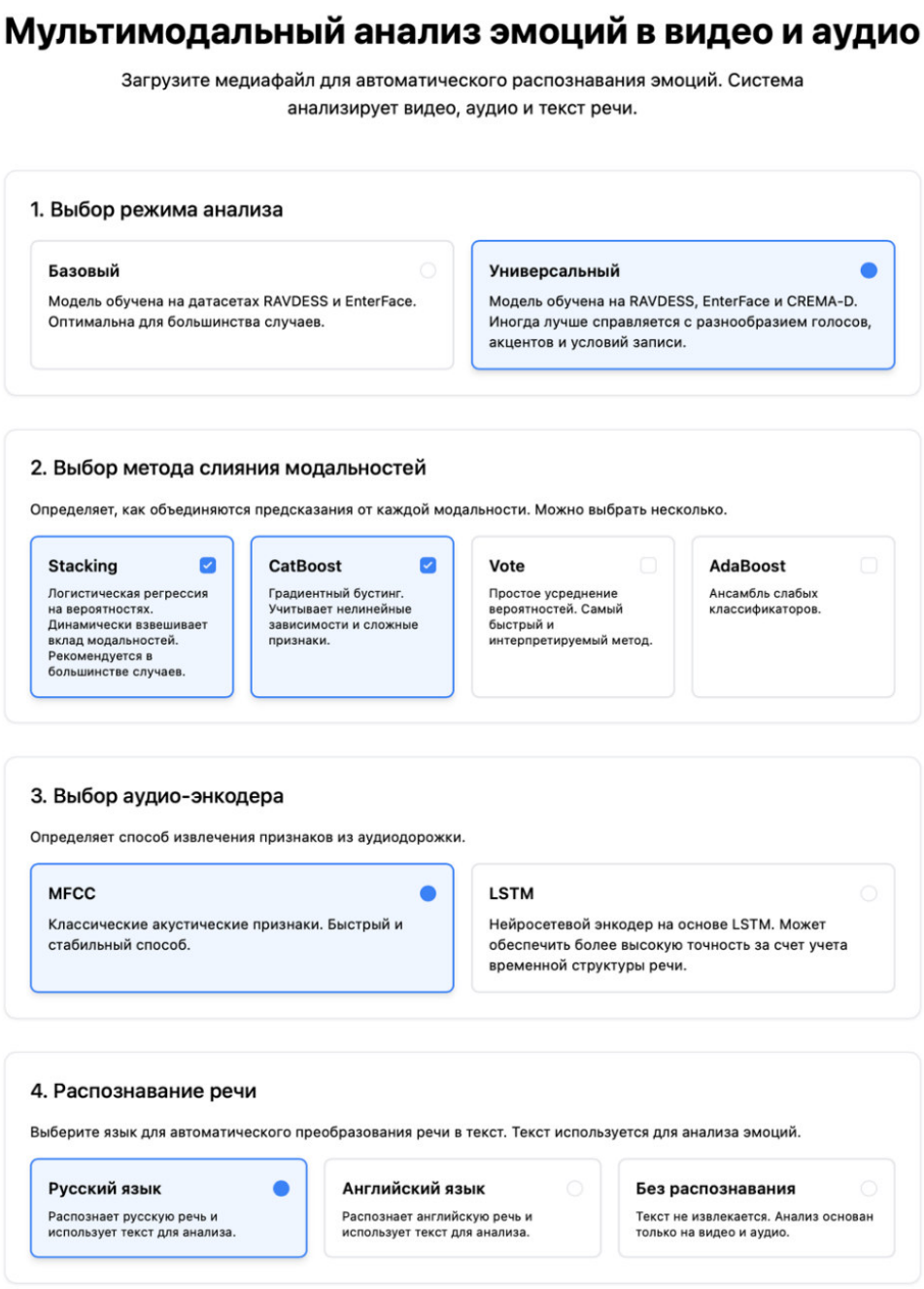

Рисунок 2. Интерфейс пользователя системы
Figure 2. System User Interface

Уровень обработки

Уровень обработки представляет собой ядро системы, именно здесь реализован основной конвейер трансформации медиаданных в аналитические признаки и итоговые предсказания. Использование технологий распределенных вычислений (Celery и Redis) позволило изолировать ресурсоемкие вычислительные операции от интерфейса управления. Процесс работы вычислительного узла включает следующие этапы.

1. Видео и аудиопотоки приводятся к единому стандарту (разрешение 720p, фиксированная частота кадров и дискретизации звука), что позволяет нейросетевым вычислительным моделям работать более стабильно.
2. Из подготовленного файла извлекается чистая аудиодорожка, которая служит основой для анализа интонаций и автоматического преобразования речи в текст. С помощью модели faster-whisper система получает текстовую расшифровку и точные временные метки для каждого произнесенного слова. Система защищена от сбоев, даже если в медиафайле отсутствует речь, система продолжит анализ по видео и аудиоканалам.

3. На основе полученных меток система сопоставляет каждый фрагмент текста с соответствующим ему отрезком аудио и набором видеокадров. Это позволяет ансамблю моделей анализировать эмоциональные проявления, зафиксированные строго в один и тот же момент времени.
4. Синхронизированные данные подаются на вход выбранной комбинации нейросетей в зависимости от настроек и выбранного метода классификации.

Данная система зарегистрирована в ФИПС как программа для ЭВМ № 2025696815 от 19.12.2025.

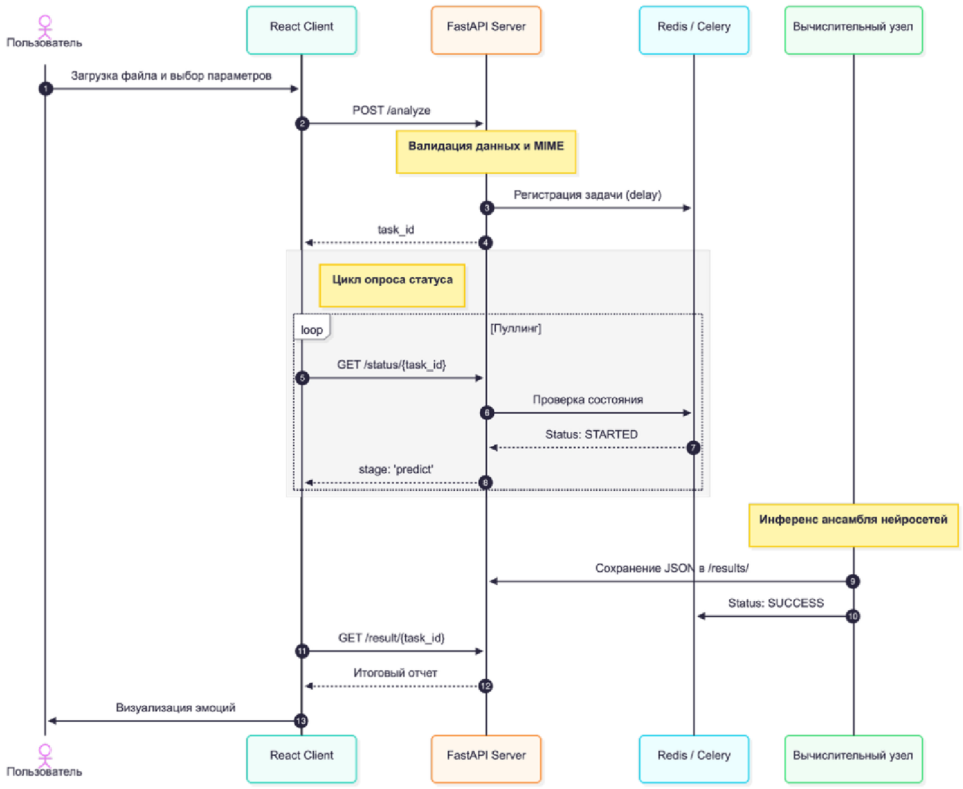

Рисунок 3. Диаграмма последовательности жизненного цикла пользовательского запроса
Figure 3. Diagram of the user query lifecycle sequence

ПРОВЕДЕНИЕ ЭКСПЕРИМЕНТА И АПРОБАЦИЯ СИСТЕМЫ

Целью экспериментального этапа являлась проверка гипотезы о том, что использование ансамблевых методов классификации в сочетании с точной временной синхронизацией позволяет повысить точность распознавания эмоций в реальных сценариях использования.

Описание набора данных

Для обучения базовых классификаторов и настройки метамоделей ансамбля использовались эталонные мультимодальные корпуса: RAVDESS, eNTERFACE и CREMA-D. Суммарный объем обучающей выборки составил более 9000 верифицированных фрагментов.

Для апробации разработанного инженерного решения в условиях, максимально приближенных к реальной эксплуатации, была сформирована специализированная выборка из 200 видеофрагментов, полученных из сервиса TikTok, на русском языке продолжительностью от 5 секунд до минуты, экспертно размеченная по семи классам эмоций и сбалансированный по гендерному (106 мужчин и 94 женщины) и возрастному составу.

Извлекаемые признаки

Процесс формирования признаков для экспериментальной выборки повторял конвейер обучения системы:

- для визуального канала – извлечение 48 кадров, нормализация и получение 2048 дескрипторов через ResNet-50;
- для акустического канала – сравнение признаков MFCC-240 и векторов AudioLSTM (128 признаков);
- для текстовой модальности – семантические эмбеддинги Multilingual BERT (768 признаков), обработанные с учетом маскирования при отсутствии речи.

Методика эксперимента

Эксперимент был направлен на оценку устойчивости разработанной системы при работе с неоднородными данными социальных медиа, характеризующимся наличием фонового шума и вариативностью освещения. Для оценки качества

распознавания использовались стандартные метрики: accuracy, precision, recall и f1-score. Особое внимание было уделено сравнительному анализу предложенного решения с существующими архитектурами мультимодального анализа:

- Tensor Fusion Network (TFN) – модель, объединяющая данные через тензорное произведение [20];
- Multi-attention Recurrent Network (MARN) – рекуррентная сеть, использующая механизмы внимания для связи видео и звука [21].

Анализ полученных результатов

В табл. 1 представлены результаты сравнительного анализа алгоритмов на специализированной выборке.

Анализ показывает, что предложенная модель на базе CatBoost с использованием мета-признаков и аудио-энкодером LSTM демонстрирует стабильное преимущество над классическими мультимодальными архитектурами TFN и MARN. Разработанный ансамбль на базе CatBoost + LSTM демонстрирует точность на 4-7% выше аналогов, что объясняется эффективным использованием показателей неопределённости (энтропии) при принятии итогового решения. Внутри системы данных подход также показал лучшие результаты, превзойдя стратегии Stacking и AdaBoost на 3-5%, а метод Voting на 8%, что показывает целесообразность использования глубокого анализа взаимосвязей между модальностями для повышения надежности классификации.

Для детального изучения сильных и слабых сторон системы была проанализирована работа лучшей модели (CatBoost + LSTM) в разрезе отдельных эмоциональных состояний в сравнении с моделью MARN. Данные по Precision, Recall и F1-score для каждого из семи классов представлены в табл. 2 и 3.

Таблица 1. Сравнение методов слияния для специализированной выборки
Table 1. Comparison of fusion methods for a specialized sample

Алгоритм	Аудио-энкодер	Accuracy	Precision	Recall	F1-score
Stacking	MFCC	0.645	0.635	0.642	0.655
Stacking	LSTM	0.660	0.651	0.658	0.669
Voting	MFCC	0.612	0.605	0.610	0.618
Voting	LSTM	0.620	0.612	0.618	0.625
AdaBoost	MFCC	0.672	0.663	0.669	0.680
AdaBoost	LSTM	0.685	0.676	0.682	0.693
CatBoost	MFCC	0.698	0.698	0.696	0.707
CatBoost	LSTM	0.715	0.705	0.712	0.720
TFN	–	0.642	0.648	0.655	0.651
MARN	–	0.678	0.682	0.688	0.685

Таблица 2. Результаты классификации по типам эмоций (модель CatBoost + LSTM)
Table 2. The results of classification by types of emotions (CatBoost + LSTM model)

Эмоция (класс)	Precision	Recall	F1-score
Удивление	0.881	0.890	0.900
Радость	0.811	0.818	0.828
Гнев	0.740	0.748	0.756
Нейтральное состояние	0.705	0.712	0.720
Отвращение	0.635	0.642	0.648
Грусть	0.599	0.605	0.611
Страх	0.564	0.569	0.577

Таблица 3. Результаты классификации по типам эмоций (модель MARN)
Table 3. Classification results by emotion type (MARN model)

Эмоция (класс)	Precision	Recall	F1-score
Удивление	0.759	0.744	0.751
Радость	0.765	0.784	0.774
Гнев	0.658	0.643	0.650
Нейтральное состояние	0.752	0.761	0.756
Отвращение	0.587	0.605	0.596
Грусть	0.679	0.697	0.688
Страх	0.574	0.582	0.578

Сравнительный анализ по типам эмоциональных состояний выявляет ключевые преимущества разработанного решения. Наиболее значительный прирост эффективности наблюдается в классе «Удивление» (+14,9%), что подтверждает высокую чувствительность предложенной связки ResNet-50 и LSTM-энкодера к динамическим изменениям мимики и интонации.

Стоит отметить, что в таких категориях, как «Нейтральное состояние», «Грусть» и «Страх», показатели моделей сопоставимы. Это объясняется спецификой эмоций, где они выражаются менее экспрессивно, что затрудняет их идентификацию даже для продвинутых архитектур. Однако за счет использования обогащенного вектора мета-признаков, предложенная система на базе CatBoost демонстрирует более высокую общую точность, особенно в сценариях с активным проявлением эмоций, где разрыв с моделью MARN становится ощутимым.

Разработанная программная система обладает широким потенциалом для внедрения в различных предметных областях. Одной из наиболее перспективных сфер применения является инжиниринг программного обеспечения, где мультимодальный сентимент-анализ может использоваться для мониторинга эмоционального состояния участников распределенных команд [22], что позволяет своевременно выявлять признаки выгорания или конфликтов. Кроме того, решение может быть интегрировано в образовательные платформы для мониторинга вовлеченности учащихся, а также в корпоративные системы поддержки принятия решений для анализа качества клиентского сервиса и эффективности внутренних коммуникаций в организациях.

ЗАКЛЮЧЕНИЕ

Проведённое исследование позволило решить актуальную инженерную задачу по созданию масштабируемой и устойчивой системы мультимодального анализа эмоциональных состояний. Разработанная программная система успешно объединяет современные методы компьютерного зрения, цифровой обработки сигналов и естественного языка в рамках единой распределенной архитектуры.

Ключевые результаты и выводы могут быть сформулированы следующим образом.

Использование трехуровневой структуры и асинхронного взаимодействия вычислительного узла под управление FastAPI, Celery и Redis обеспечило необходимую производительность для обработки медиаконтента в условиях реального времени. Модульность системы позволяет гибко настраивать конвейер обработки данных, включая выбор аудиоэнкодера, метода слияния и выбор языка речи.

Реализованный механизм временной синхронизации модальностей на основе временных меток обеспечил точность сопоставление мимики, интонации и семантики.

Основным результатом стала стратегия поздней интеграции (Late Fusion) на базе мета-классификатора CatBoost. Обучение модели на обогащенном векторе признаков, включающем метрики неопределенности (энтропию и маржу вероятностей), позволяет нивелировать ошибки отдельных каналов при работе с «шумными» данными.

В ходе апробации на специализированной выборке, предложенный ансамбль продемонстрировал превосходство над классическими архитектурами TFN и MARN, а также другими методами ансамблирования: Stacking, Voting, AdaBoost, достигнув точности (Accuracy) = 0,715 и F1-меры = 0,720.

Представленное инженерное решение является готовым инструментом для автоматизации мониторинга эмоционального состояния. Направления дальнейших исследований связаны с внедрением механизмов внимания для более точного сопоставления временных интервалов различных модальностей.

Авторы заявляют об отсутствии конфликта интересов.
The authors declare no conflict of interest.

ЛИТЕРАТУРА

1. **Орлов А. А., Миронов М. И., Абрамова Е. С.** Обзор и анализ подходов и практических областей применения распознавания эмоций человека. *Вестник ЮУрГУ. Серия «Компьютерные технологии, управление, радиоэлектроника»*. 2023. Т. 23. № 4. С. 5–15. DOI: 10.14529/ctcr230401.
2. **Кутузова А. С., Чернов Н. А., Котенев Т. Е.** Применение методов искусственного интеллекта в разработке алгоритма оценки усталости водителя. *Известия высших учебных заведений. Серия «Экономика, финансы и управление производством» [Ивэкофин]*. 2025. № 4(66). С. 88-95. DOI: 10.6060/ivecofin.2025664.748.
3. **Cassee N., Agaronian A., Constantinou E., Novielli N., Serebrenik A**. Transformers and meta-tokenization in sentiment analysis for software engineering. *Empirical Software Engineering*. 2024. Vol. 29. N 77. DOI: 10.1007/s10664-024-10468-2.

REFERENCES

1. **Orlov A.A, Mironov M.I, Abramova E.S.** Review and analysis of approaches and practical applications of human emotion recognition. *Bulletin of the South Ural State University. Ser. Computer Technologies, Automatic Control, Radio Electronics*. 2023 N 23(4). P. 5–15. DOI: 10.14529/ctcr230401. (in Russian).
2. **Kutuzova A.S., Chernov N.A., Kotenev T.E.** Using artificial intelligence methods in the development of an algorithm for assessing driver fatigue. *Ivecofin*. 2025. N 04(66). P. 88-95. DOI: 10.6060/ivecofin. 2025664.748. (in Russian).
3. **Cassee N., Agaronian A., Constantinou E., Novielli N., Serebrenik A.** Transformers and meta-tokenization in sentiment analysis for software engineering. *Empirical Software Engineering*. 2024. Vol. 29. N 77. DOI: 10.1007/s10664-024-10468-2.

4. **Загидуллина М. В**. Современное состояние мультимодального анализа: к вопросу о перспективах метода. *Научный результат. Социальные и гуманитарные исследования*. 2023. Т. 9. № 1. С. 84–99. DOI: 10.18413/2408-932X-2023-9-1-0-7.
5. **Poria S., Cambria E., Bajpai R., Hussain A.** A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*. 2017. Vol. 37. P. 98–125. DOI: 10.1016/j.inffus.2017.02.003.
6. **Baltrusaitis T., Ahuja C., Morency L**. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2019. Vol. 41. N 2. P. 423–443. DOI: 10.1109/TPAMI.2018.2798607.
7. **Дорохин М.А., Чернышев С.А**. Нейросетевая диагностика произношения на английском языке. Разработка вебсервиса. *Программные продукты и системы*. 2025. Т. 38. № 4. С. 724–732. DOI: 10.15827/0236-235X.152.724-732.
8. **Антипова С.А**. Разработка системы контроля доступа на основе распознавания лиц. *Программные продукты и системы*. 2021. Т. 34. № 2. С. 245–256. DOI: 10.15827/0236-235X.134.245-256.
9. **He K., Zhang X., Ren S., Sun J.** Deep Residual Learning for Image Recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 2016. P. 770–778. DOI: 10.1109/CVPR.2016.90-235X.134.245-256.
10. **Greff K., Srivastava R.K., Koutník J., Steunebrink B.R., Schmidhuber J.** LSTM: A Search Space Odyssey. *IEEE Transactions on Neural Networks and Learning Systems*. 2017. Vol. 28. N 10. P. 2222–2232. DOI: 10.1109/TNNLS.2016.2582924.
11. **Radford A., Kim J.W., Xu T., Brockman G., McLeavey C., Sutskever I.** Robust Speech Recognition via Large-Scale Weak Supervision. *Proceedings of the 40th International Conference on Machine Learning*. 2023. Vol. 202. P. 28448–28481. DOI: 10.48550/arXiv.2212.04356.
12. **Devlin J., Chang M., Lee K., Toutanova K.** BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL*. 2019. P. 4171–4186. DOI: 10.18553/v1/N19-1423.
13. **Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V.** Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 2011. Vol. 12. P. 2825–2830. DOI: 10.48550/arXiv.1201.0490.
14. **Ganaie M.A., Hu M., Malik A.K., Tanveer M., Suganthan P.N.** Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*. 2022. Vol. 115. P. 105151. DOI: 10.1016/j.engappai.2022.105151.
15. **Prokhorenkova L., Gusev G., Vorobev A., Dorogush A.V., Gulin A.** CatBoost: unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*. 2018. Vol. 31. DOI: 10.48550/arXiv.1706.09516.
16. **Фазульянов Д.В., Гусева А.И.** Разработка мультимодального метода сентимент-анализа для поддержки принятия решений в организациях. *Современные наукоемкие технологии. Региональное приложение*. 2024. № 5-2. С. 313–320. DOI: 10.17513/snt.40045.
17. **Livingstone S.R., Russo F.A.** The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). *PLoS ONE*. 2018. Vol. 13. N 5. DOI: 10.1371/journal.pone.0196391.
18. **Martin O., Kotsia I., Macq B., Pitas I.** The eNTERFACE'05 Audio-Visual Emotion Database. *IEEE International Conference on Multimedia and Expo (ICME).* 2006. P. 8–8. DOI: 10.1109/ICDEW.2006.145.
19. **Cao H., Cooper D.G., Keutmann M.K., Gur R.C., Nenkova A., Verma R**. CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset. *IEEE Transactions on Affective Computing*. 2014. Vol. 5. N 4. P. 377–390. DOI: 10.1109/TAFFC.2014.2336244.

4. **Zagidullina M.V.** The current state of multimodal analysis: on the question of the prospects of the method. *Research Result. Social Studies and Humanities.* 2023. N 9(1). P. 84–99. DOI: 10.18413/2408-932X-2023-9-1-0-7. (in Russian).
5. **Poria S., Cambria E., Bajpai R., Hussain A.** A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*. 2017. Vol. 37. P. 98–125. DOI: 10.1016/j.inffus.2017.02.003.
6. **Poria S., Cambria E., Bajpai R., Hussain A.** A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*. 2017. Vol. 37. P. 98–125. DOI: 10.1109/TPAMI.2018.2798607.
7. **Dorokhin M.A, Chernyshev S.A**. Neural network-based pronunciation diagnostics for English. Web service development. *Software Products and Systems*. 2025. Vol. 38. N 4. P. 724–732. DOI: 10.15827/0236-235X.152.724-732. (in Russian).
8. **Antipova S.A.** The access control system development based on face recognition. *Software & Systems*. 2021. Vol. 34. N 2. P. 245–256. DOI: 10.15827/0236-235X.134.245-256. (in Russian).
9. **He K., Zhang X., Ren S., Sun J.** Deep Residual Learning for Image Recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 2016. P. 770–778. DOI: 10.1109/CVPR.2016.90.
10. **Greff K., Srivastava R.K., Koutník J., Steunebrink B.R., Schmidhuber J.** LSTM: A Search Space Odyssey. *IEEE Transactions on Neural Networks and Learning Systems*. 2017. Vol. 28. N 10. P. 2222–2232. DOI: 10.1109/TNNLS.2016.2582924.
11. **Radford A., Kim J.W., Xu T., Brockman G., McLeavey C., Sutskever I.** Robust Speech Recognition via Large-Scale Weak Supervision. *Proceedings of the 40th International Conference on Machine Learning*. 2023. Vol. 202. P. 28448–28481. DOI: 10.48550/arXiv.2212.04356.
12. **Devlin J., Chang M., Lee K., Toutanova K.** BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL*. 2019. P. 4171–4186. DOI: 10.18553/v1/N19-1423.
13. **Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V.** Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 2011. Vol. 12. P. 2825–2830. DOI: 10.48550/arXiv.1201.0490.
14. **Ganaie M.A., Hu M., Malik A.K., Tanveer M., Suganthan P.N.** Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*. 2022. Vol. 115. P. 105151. DOI: 10.1016/j.engappai.2022.105151.
15. **Prokhorenkova L., Gusev G., Vorobev A., Dorogush A.V., Gulin A.** CatBoost: unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*. 2018. Vol. 31. DOI: 10.48550/arXiv.1706.09516.
16. **Fazulianov D.V, Guseva A.I.** Development of a multimodal method of sentiment analysis to support decision-making in organizations. *Modern High Technologies. Regional application*. 2024. N (5-2). P. 313–320. DOI: 10.17513/snt.40045. (in Russian).
17. **Livingstone S.R., Russo F.A.** The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). *PLoS ONE*. 2018. Vol. 13. N 5. DOI: 10.1371/journal.pone.0196391.
18. **Martin O., Kotsia I., Macq B., Pitas I.** The eNTERFACE'05 Audio-Visual Emotion Database. *IEEE International Conference on Multimedia and Expo (ICME).* 2006. P. 8–8. DOI: 10.1109/ICDEW.2006.145.
19. **Cao H., Cooper D.G., Keutmann M.K., Gur R.C., Nenkova A., Verma R**. CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset. *IEEE Transactions on Affective Computing*. 2014. Vol. 5. N 4. P. 377–390. DOI: 10.1109/TAFFC.2014.2336244.

20. **Zadeh A., Chen M., Poria S., Cambria E., Morency L.P.** Tensor Fusion Network for Multimodal Sentiment Analysis. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP).* 2017. P. 1103–1114. DOI: 10.18553/v1/D17-1115.
21. **Zadeh A., Liang P.P., Poria S., Vij P., Cambria E., Morency L.P.** Multi-attention Recurrent Network for Human Communication Comprehension. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2018. Vol. 32. N 1. P. 5642–5649. DOI: 10.1609/aaai.v32i1.12011.
22. **Зимнуров М.Ф**. Разработка интерфейса по метрике загруженности сотрудников. *Известия высших учебных заведений. Серия «Экономика, финансы и управление производством» [Ивэкофин]*. 2024. № 4(62). С. 82–87. DOI: 10.6060/ivecofin.2024624.705.

20. **Zadeh A., Chen M., Poria S., Cambria E., Morency L.P.** Tensor Fusion Network for Multimodal Sentiment Analysis. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP).* 2017. P. 1103–1114. DOI: 10.18553/v1/D17-1115.
21. **Zadeh A., Liang P.P., Poria S., Vij P., Cambria E., Morency L.P.** Multi-attention Recurrent Network for Human Communication Comprehension. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2018. Vol. 32. N 1. P. 5642–5649. DOI: 10.1609/aaai.v32i1.12011.
22. **Zimnurov M.F.** Developing an interface based on employee workload metrics. *Ivecofin*. 2024. N 4 (62). P. 82–87. DOI: 10.6060/ivecofin.2024624.705. (in Russian).

Поступила в редакцию 24.01.2026
Принята к опубликованию 07.02.2026

Received 24.01.2026
Accepted 07.02.2026