ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ОБРАЗОВАНИЯ
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

**Факультет информатики, математики и компьютерных наук**

**Программа подготовки бакалавров по направлению**
**01.03.02 Прикладная математика и информатика**

# ОТЧЕТ

## по проектной работе

Разработка многомодальной системы распознавания эмоциональных состояний человека

**Выполнил студент гр.** 20ПМИ-2
Сидорова Анна Павловна

**Руководитель проекта:**
специалист по анализу данных
Абросимов Кирилл Игоревич

**Оценка**: 10

21.03.2023

Нижний Новгород, 2023

**Содержание**

# 1 Общее описание проекта

## 1.1 Описание компании

**Заказчик** - Национальный исследовательский университет «Высшая школа экономики».

**Инициатор** - Factory5 Dev (ООО «M5»).

**Руководитель проекта** - специалист по анализу данных Абросимов Кирилл Игоревич.

**Место работы по проекту** - дистанционно.

## 1.2 Описание технического задания

**Цель**: система, распознающая эмоциональное состояние человека по видеомодальности, аудиомодальности и текстовой модальности.

**Задачи**:

а) Изучить современные (state-of-the-art) статьи по задаче распознавания эмоций;

б) Изучить и произвести статистический анализ открытого набора данных MELD;

в) Изучить и применить методы классического обучения;

г) Изучить и применить современные архитектуры нейронных сетей;

д) Изучить и применить методы извлечения признаков из видеомадальности (сверточные нейронные сети);

е) Изучить и применить методы извлечения признаков из аудиомодальности (mel-спектрограммы, toolkit openSMILE);

ж) Изучить и применить методы извлечения признаков из текстовой модальности (Вектора, на основе тональных словарей, BERT);

з) Для каждой модальности построить классификатор и оценить качество;

и) Объединить модальности на уровне векторов и построить общий классификатор, оценить качество;

к) Объединить модальности на уровне принятия решений, т.е. Построить классификатор, принимающий решение на основе предсказаний

локальных классификаторов, построенных на определенной модальности.

# 2 Содержательная часть

## 2.1 Ход работы

Моя работа состояла из следующих шагов:

**Содержание**



### 2.1.1 Распределение задач

В начале работ было выдвинуто три параллельные задачи - создание унимодальной модели для текста, аудио и видео. Так как изначально участников проекта было трое, каждый взял по одной модальности. Мой выбор пал на аудио.

### 2.1.2 Общая часть

Командно был выполнен анализ сета данных и также было изучено 6 статей, работа которых была основана на датасете MELD (тот же сет данных, что и требовался по ТЗ).

Краткая выжимка из анализа:

| Статистика | Тренировочная | Валидационная | Тестовая |
|---|---|---|---|
| Модальности | {а, в, т} | {а, в, т} | {а, в, т} |
| # уникальных слов | 5783 | 1486 | 2529 |
| Средняя длина фраз | 8.63 | 8.61 | 8.90 |
| Макс длина фраз | 72 | 42 | 50 |
| Сред. # эмоций диалог | 3.30 | 3.35 | 3.24 |
| # диалогов | 1038 | 114 | 280 |
| # фраз | 9989 | 1108 | 2610 |
| # говорящих | 260 | 47 | 100 |
| # смен эмоций | 5358 | 605 | 1353 |
| Сред. длина фраз | 3.16s | 3.14s | 3.12s |

Распределение эмоций в MELD:

| Эмоции | Тренировочная | Валидационная | Тестовая |
|---|---|---|---|
| Злость | 1109 | 153 | 345 |
| Отвращение | 271 | 22 | 68 |
| Страх | 268 | 40 | 50 |
| Радость | 1743 | 163 | 402 |
| Нейтральное | 4710 | 469 | 1256 |
| Грусть | 683 | 111 | 208 |
| Удивление | 1205 | 150 | 281 |

**Минусы датасета**:

а) Датасет несбалансированный. Класс нейтрального настроения составляет 48% от всей выборки;

б) Опираясь на описание из файлов .csv, некоторые видео не соответствуют фразе/диалогу/высказыванию, что будет путать модели в будущем;

в) В одном диалоге могут присутствовать много актеров;

г) Много посторонних звуков.

**Плюсы работы с датасетом**:

а) Датасет действительно большой. Даже отбросив “поломанные” данные можно найти немало наблюдений для построения моделей;

б) Справиться с вышеуказанными минусами тяжело, так что работу с датасетом можно назвать вызовом.

### 2.1.3 Построение унимодальной модели для аудиоданных

В данной секции будет рассмотрены результаты исследования моделей для аудио модальности.

Параллельно Екатериной Дудниковой изучались модели для видео модальности.

#### 2.1.3.1 Подготовка аудиоданных

Изначально аудиоданные в рассматриваемом датасете отсутствуют. Следовательно, сперва предстояло их выделить с помощью модуля moviepy. Аудио были названы в том же стиле, что и видео вида diaX_uttY.mp3.

#### 2.1.3.2 Анализ

Далее была проанализирована длина аудио отрывков:

| Модуль | # аудио | Средняя длина | Макс длина | Мин длина |
|---|---|---|---|---|
| Обучающая | 9989 | 3.2 | 41.1 | 0.13 |
| Валидационная | 1108 | 3.17 | 28.6 | 0.13 |
| Тестовая | 2610 | 3.17 | 16.8 | 0.18 |

#### 2.1.3.3 Исследуемые работы признаков

Всего было отобрано 3 набора признаков. Два набора по ТЗ - мел-спектрограммы и openSMILE. Третий же - мел-кепстральные коэф-

фициенты - был взят из-за плохих результатов модели на мел-спектрограммах.

Определение описанных выше признаков раскрыто в прил. А.

#### 2.1.3.4 Предобработка

**Для мел-спектрограмм и мел-кепстральных коэффициентов:**

Данные, а именно сигнал, надо привести к одному размеру для обучения и предсказания. Размер - средняя длина аудио записей из train выборки.

Длина сигнала высчитывается по формуле частота · средняя длина. Сигналы, которые больше полученной длины, обрезаются. Иные же трансформируются отзеркаливанием до обозначенной длины, т.е. запись типа $[1, 2, 3]$ при длине 7 становится $[1, 2, 3, 3, 2, 1, 1]$. Это сделано для уменьшения вероятности появления артефактов.

После обработки длины аудио был выполнен пре-эмфазис. Для начала разберемся что это такое.

**Пре-эмфазис** - распространенный инструмент предварительной обработки, используемый для компенсации средней формы спектра, который подчеркивает более высокие частоты.

Выражается следующим образом:

$$y_n = x_n - \alpha x_{n-1},$$

где $y_n$ - отфильтрованный сигнал, полученный из разницы амплитуд $x_n$ и $\alpha x_{n-1}$. Для $\alpha$ обычно значение берется 0.97.

**Примеры:**

Почему “укорачивание“ не грозит потерей эмоций? Вероятность того, что некоторые данные будут потеряны, существует. Однако, есть теория, что для распознавания эмоций достаточно некоторого отрывка, потому что если человек показывает грусть на протяжении минуты, то и в любой момент этой минуты он будет грустным.

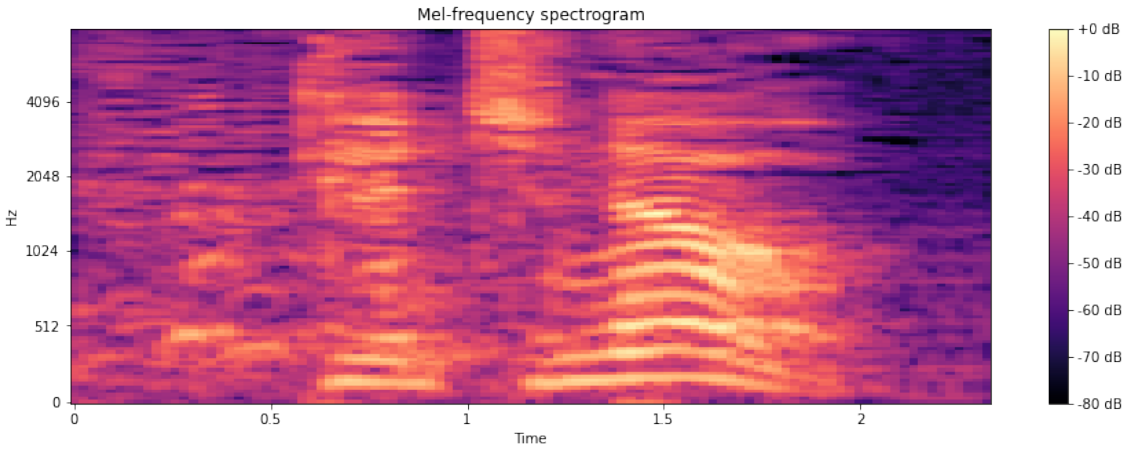


Рисунок 2.1 — До преобразований

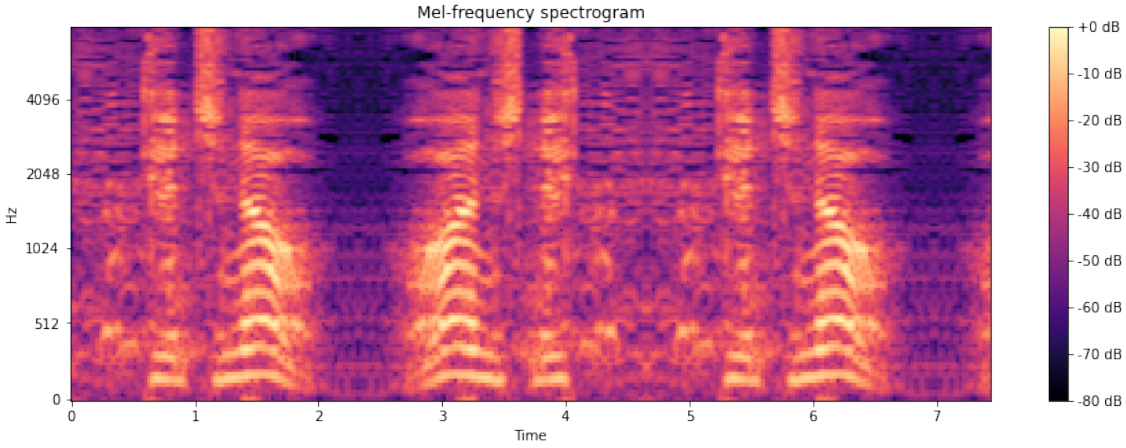


Рисунок 2.2 — Обрезка + отзеркаливание

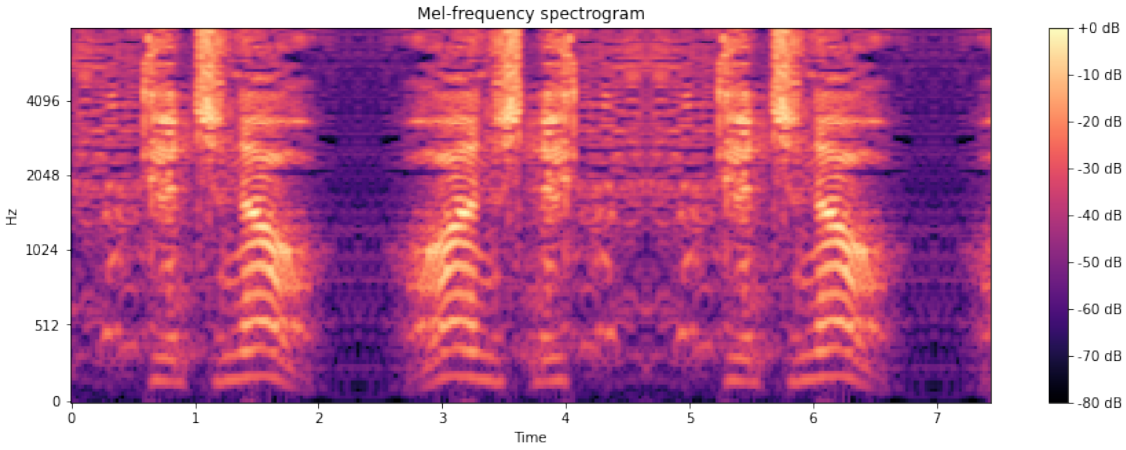


Рисунок 2.3 — Пре-эмфазис

**Для openSMILE:** OpenSMILE имеет два режима:

а) набор признаков, характеризующих весь аудио-отрывок;

б) набор векторов, где каждый вектор обрабатывает некоторое окно.

В работе был использован первый режим, поэтому обрезать аудио не было смысла. Пре-эмфазис тоже не требовался, так как openSMILE сам его использует при формировании признаков.

Таким образом, toolkit openSMILE берет “сырую“ аудиозапись и обрабатывает её.

#### 2.1.3.5 Исследование моделей

Для интерпретации матриц ошибок представлены следующие значения классов:

| Класс | Значение |
|---|---|
| anger | 0 |
| disgust | 1 |
| fear | 2 |
| sadness | 3 |
| neutral | 4 |
| joy | 5 |
| surprise | 6 |

Для аудио были опробованы следующие модели:

а) MLP для OpenSMILE;
б) SVM для OpenSMILE;
в) CNN для мел-спектрограмм;
г) CNN для мел-кепстральных коэффициентов.

В итоге это дало следующие результаты на test’ой выборке:

| Модель | balanced_accuracy_score | accuracy_score | f1 weighted |
|---|---|---|---|
| MLP | 0.202614 | 0.369349 | 0.37526 |
| SVM | 0.264384 | 0.360536 | 0.376306 |
| CNN мс | 0.1429 | 0.4812 | 0.35 |
| CNN MFCC | 0.1429 | 0.4812 | 0.35 |

#### 2.1.3.6 Выбор и описание итоговой модели для аудио

По результатам видно, что итоговой моделью был выбран метод опорных векторов. Основным набором признаков стал openSMILE.

Перед тем, как искать модель для openSMILE, было решено из-за большой размерности применить метод главных компонент.

Метод главных компонент — один из основных способов уменьшить размерность данных, потеряв наименьшее количество информации. Позволяет разделить матрицу исходных данных X на две части: «содержательную» и «шум».

Было выставлено количество компонент равным 2000, потеря информации составила 0.03%.

**Результаты**:

| Оценка | train | test |
|---|---|---|
| balanced_accuracy_score | 0.793905 | 0.264384 |
| accuracy_score | 0.660721 | 0.360536 |
| f1 weighted | 0.662948 | 0.376306 |

**Confusion matrix:**

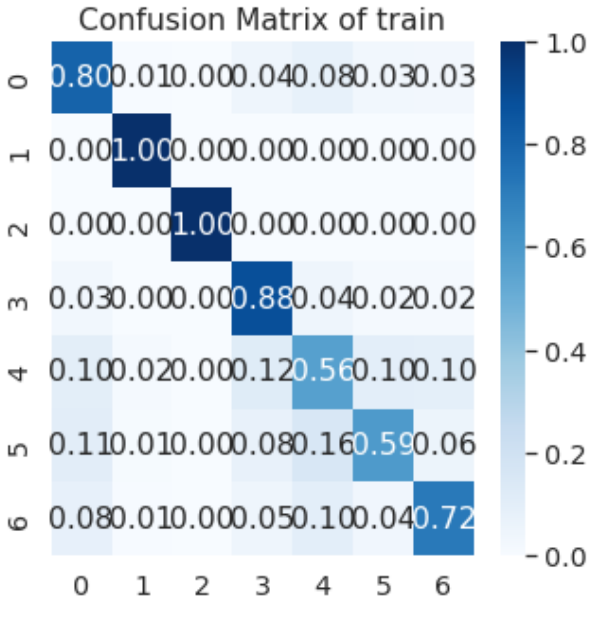


Рисунок 2.4 — В обучающей выборке

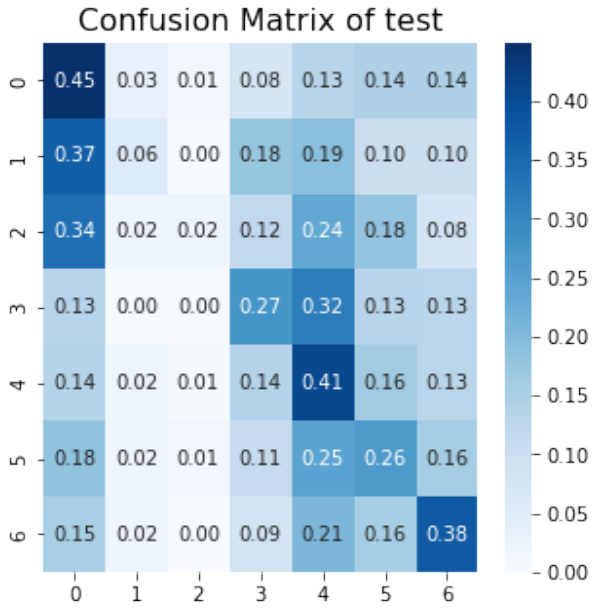


Рисунок 2.5 — В тестовой выборке

### 2.1.4 Построение унимодальной модели для текстовых данных

На момент построение аудио модели было сообщено об уходе одного из участников проекта, отвечавшего за текстовую модальность, в академический отпуск.

Было предложено строить модель не из трёх модулей (текст, аудио, видео), а из двух (аудио, видео). Однако текстовая модальность предположительно давала бы самый высокий результат из других модальностей, поэтому было решено найти быстрое решение данной проблемы, но при этом не сильно терявшее в метриках. Таким образом был выбран трансформер RoBERTa, так как он хорошо справляется с задачей классификации эмоций и был задействован во многих проанализированных статьях.

Параллельно с этим Екатерина Дудникова исследовала классические методы NLP для текстовой модальности.

#### 2.1.4.1 Предобработка

Текст предобработки не проходил, так как был задействован токенайзер RoBERT'ы.

Предобработка перед токенайзером не нужна, так как он уже имеет свой словарь и сам предоставляет лексический сканер.

#### 2.1.4.2 Описание итоговой модели для текста

**Полученные результаты:**

| Оценка | train | test |
|---|---|---|
| balanced_accuracy_score | 0.512 | 0.399841 |
| accuracy_score | 0.763 | 0.653257 |
| f1 weighted | 0.74 | 0.627947 |

**Confusion matrix:**

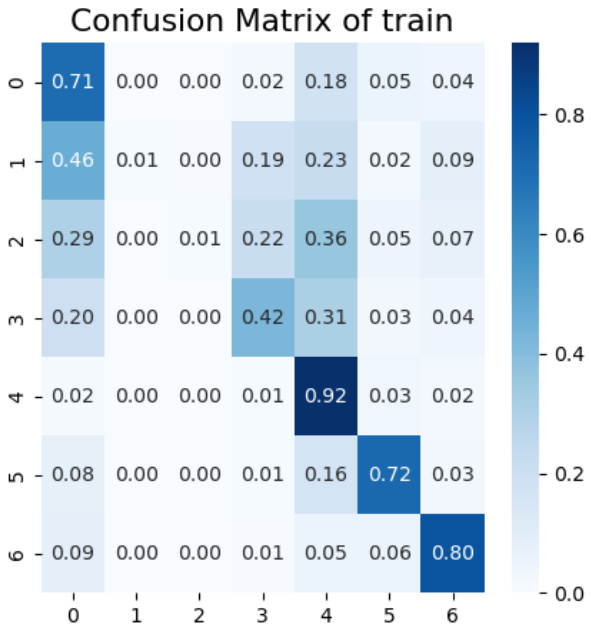


Рисунок 2.6 — В обучающей выборке

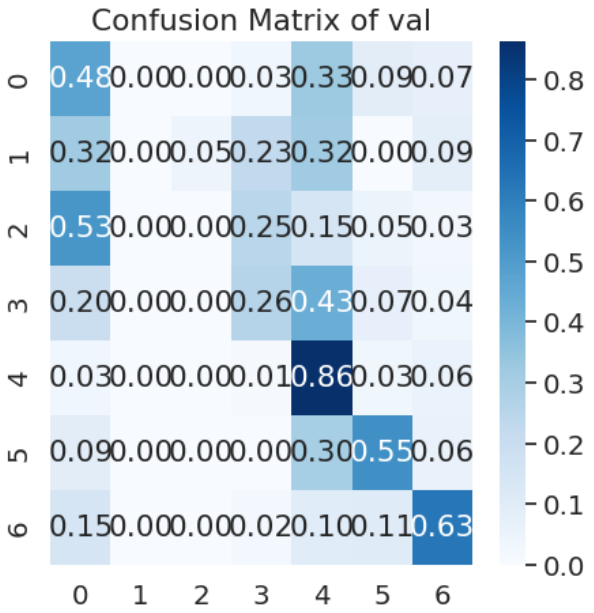


Рисунок 2.7 — В валидационной выборке

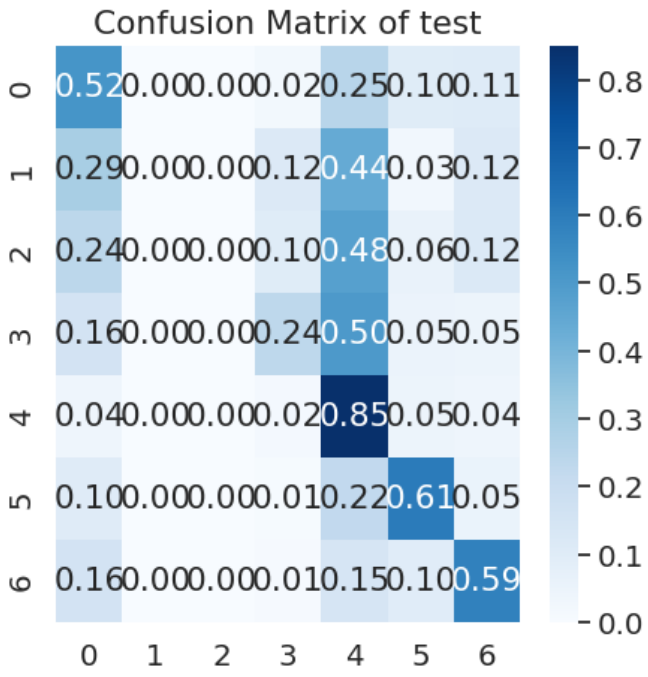


Рисунок 2.8 — В тестовой выборке

**Минусы:**

а) Скорее всего сложность в обучении связана наличием большого кол-ва коротких слов, имеющих несколько значений (What?, Oh! etc);

б) Все же для текста эмоции отвращения и страха являются сложными. Модель не смогла выучить и предсказать данные классы в достаточной мере.

**Плюсы:**

а) Результат для сложной задачи предсказания эмоций является неплохим. По сравнению с предыдущими модальностями он дал высокие результаты.

### 2.1.5 Подытоживание к унимодальным системам

В итоге RoBERTa (текст) получила лучшую оценку, чем классические методы NLP (текст).

Если сравнивать с текущим наилучшим результатом среди зафиксированных простроенных унимодальностей (в статье [1]) по метрике accuracy, то можно наблюдать следующую картину:

| Модель only Audio | текущая | статья [1] |
|---|---|---|
| accuracy_score | 0.360536 | 0.4904 |
| f1 weighted | 0.376306 | 0.3963 |

| Модель only Text | текущая | статья [1] |
|---|---|---|
| accuracy_score | 0.653257 | 0.6724 |
| f1 weighted | 0.627947 | 0.6623 |

Метрики accuracy score и f1 weighted не являются показательными в данной задаче, так как их увеличение не гарантирует улучшение обобщаемости модели (а больше недообучение в класс neutral, что показано в аудио модели на основе признаков мел-спектрограмм). Поэтому нельзя точно сказать, насколько модели из статей являются эталонными, так как лучших метрих (например, balanced accuracy) и матриц ошибок показано не было.

### 2.1.6 Многомодальная система

На момент начала исследований многомодальных систем участников проекта осталось двое, что совпадало с количеством исследуемых

моделей. Для изучения и построения были представлены early fusion и late fusion. Мною был выбран early fusion, о нём и пойдет речь далее.

Параллельно с моей работой проводилось изучение и разработка модели late fusion Екатериной Дудниковой.

Идейно, early fusion интегрирует необработанные данные из отдельных модальностей в унифицированное представление перед продолжением процесса обучения/извлечения признаков. Проще говоря - берутся лучшие признаки из каждой модальности, конкатенируются и на основе полученных векторов обучаются.

Таким образом были получены эмбеддинги из видео (число компонент 512) и текста (число компонент 768), а так же openSMILE из аудио (число компонент 6373). Далее все вектора признаков были уменьшены в размерности с помощью метода главных компонент. Каждый вектор был размером 1 на 512, итоговый 1 на 1536.

Исходя из проделанных действий, были построены два классификатора и также были подобраны коэффициенты. Результаты вы можете наблюдать в таблице ниже:

а) SVM C=1

| SVM | train | test |
|---|---|---|
| balanced_accuracy_score | 0.667671 | 0.467386 |
| accuracy_score | 0.752095 | 0.615709 |
| f1 weighted | 0.761351 | 0.625343 |

б) SVM C - подобранный параметр

| SVM | train | test |
|---|---|---|
| balanced_accuracy_score | 0.692083 | 0.469572 |
| accuracy_score | 0.762999 | 0.61341 |
| f1 weighted | 0.770736 | 0.622541 |

**Confusion matrix:**

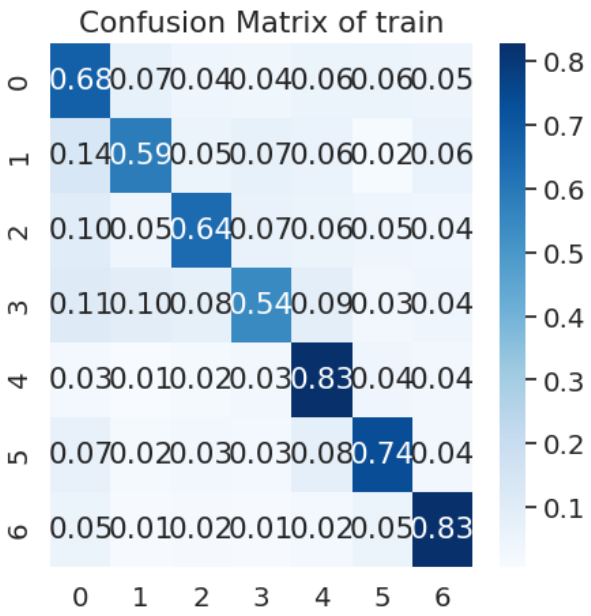


Рисунок 2.9 — В обучающей выборке

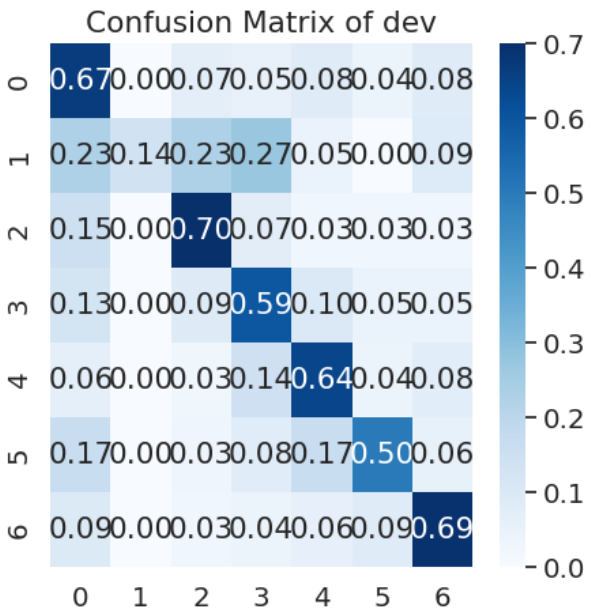


Рисунок 2.10 — В валидационной выборке

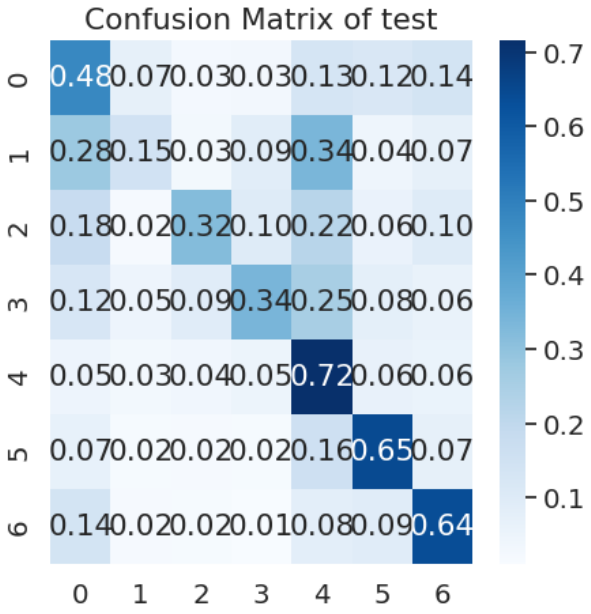


Рисунок 2.11 — В тестовой выборке

Так же были проведены эксперименты с размерностью вектора текстового эмбеддинга. С помощью полносвязного слоя размерность была уменьшена с 768 до 512. Результаты:

| SVM | train | test |
|---|---|---|
| balanced_accuracy_score | 0.65685 | 0.459055 |
| accuracy_score | 0.744435 | 0.616475 |
| f1 weighted | 0.755048 | 0.6250211 |

## 2.2 Описание результатов проекта

| Результаты | balanced_accuracy_score | accuracy_score | f1 weighted |
|---|---|---|---|
| Only Video | 0.156249 | 0.37445 | 0.30669 |
| Only Audio | 0.264384 | 0.360536 | 0.376306 |
| Only Text | 0.399841 | 0.653257 | 0.627947 |
| Late Fusion | 0.413862 | 0.6341 | 0.622822 |
| Early Fusion | 0.469572 | 0.61341 | 0.622541 |

В сравнении с указанной статьей [1]:

| Результаты | accuracy | f1 weighted | accuracy статья | f1 weighted статья |
|---|---|---|---|---|
| Only Video | 0.37445 | 0.30669 | 0.4563 | 0.3244 |
| Only Audio | 0.360536 | 0.376306 | 0.4904 | 0.3963 |
| Only Text | 0.653257 | 0.627947 | 0.6724 | 0.6623 |
| Late Fusion | 0.6341 | 0.622822 | 0.6785 | 0.6671 |
| Early Fusion | 0.61341 | 0.622541 | 0.6728 | 0.6681 |

Итоговым продуктом проекта стал бот в Telegram и локальная библиотека. Ссылку на бота вы найдете в прил. Б.

Моей ролью в итоговом продукте было написание локальной библиотеки и использование предобученной трансформерной нейронной сети wav2vec для автоматического распознавания речи в текст (нс wav2vec2-base-960h). Нейронная сеть STT (speech-to-text) была нужна для того, чтобы получать из видео все необходимые модальности. В данном случае это текст, аудио было взято с помощью библиотеки moviepy из видео. Таким образом на вход принимается только видеоролик.

## 2.3 Описание использованных в проекте способов и технологий

Ниже приведены использованные мною технологии:

а) Признаки аудио модальности (мел-спектрограммы, openSMILE, MFCC);

б) Классический ML (метод опорных векторов, метод главных компонент);

в) Нейронные сети;

г) Трансформеры (RoBERTa, wav2vec);

д) Early fusion;

е) Прочее (программирование на Python, использование модулей numpy, pandas, Jupyter notebooks, pytorch, torchaudio, librosa, transformers, opensmile, pickle ect).

## 2.4 Описание своей роли в проектной команде

В главах выше подробно были описаны модули над которыми я работала. Поэтому кратко просуммирую все ниже:

а) Унимодальная модель для аудио;

б) Унимодальная модель (трансформеры) для текста;

в) Объединение модальностей методом раннего слияния;

г) Локальная библиотека для Telegram бота (Python);

д) Трансформер STT.

## 2.5 Описание отклонений и трудностей, возникших в ходе выполнения проекта

Во время хода проекта я столкнулась со следующими трудностями:

а) Во время начала проекта не было знаний о работе с аудио данными;

б) Не было знаний о нейронных сетях и трансформерах;

в) Очень не сбалансированный датасет, нужна была новая метрика, которая качественно бы отображала интерпретируемость полученных моделей;

г) Много теории и исследований, поэтому порой дедлайны были просрочены;

д) Уход участника проекта, а в следствии чего его часть работ была переложена на нас;

е) Неверное указание сроков со стороны ВШЭ, сроки в середине работы были резко сокращены;

ж) Небольшие трудности в работе с напарницей из-за разного отношения к дедлайнам;

з) Не успела подготовить библиотеку для загрузки в open source.

# 3 Заключение

Как по мне работа была проделана огромная, было получено много новых знаний. Прийдя на этот проект у нас была лишь база ML, поэтому за время проекта у меня получилось:

а) Научиться работать с аудио данными;

б) Узнать о новых признаках, наборах инструментов;

в) Узнать как работать с нейросетями и предобученными моделями;

г) Побороть страх перед защитой проекта;

д) Научиться работать с большим объемом информации на английском языке;

е) Узнать о методах слияния моделей (ансамблирование, работа на уровне принятия решений, раннее слияние);

ж) Узнать о новых ML задачах, таких как STT;

з) Придумать новые идеи для дальнейшей реализации.

Удалось придумать следующие идеи для улучшения интерпритируемости модели:

а) Использовать в будущем более сложные нейронные сети на мел-спектрограммах;

б) Добавить к набору признаков мел-спектрограмм мел-спектрограммы гармоничных и ударных звуков;

в) Использовать другой датасет, где будет иммитация разговора человека с компьютером (MOSEI, например).

Код проекта был приложен в архиве при загрузке в LMS.

Из соображений конфиденциальности показаны в коде только лучшие модели.

# Список использованных источников

1. V. Chudasama, P. Kar, A. Gudmalwar, N. Shah, P. Wasnik, and N. Onoe. *M2FNet: Multi-modal Fusion Network for Emotion Recognition in Conversation.* In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pages 4652–4661, June 2022.

# Приложение А Теоретическая справка

Определения, которые могут понадобиться:

а) Спектр - среднее статистическое значение определенного сигнала или типа сигнала (включая шум), проанализированное с точки зрения его частотного содержания.
Например, форма звуковой волны с течением времени (слева) имеет широкий спектр мощности звука (справа):

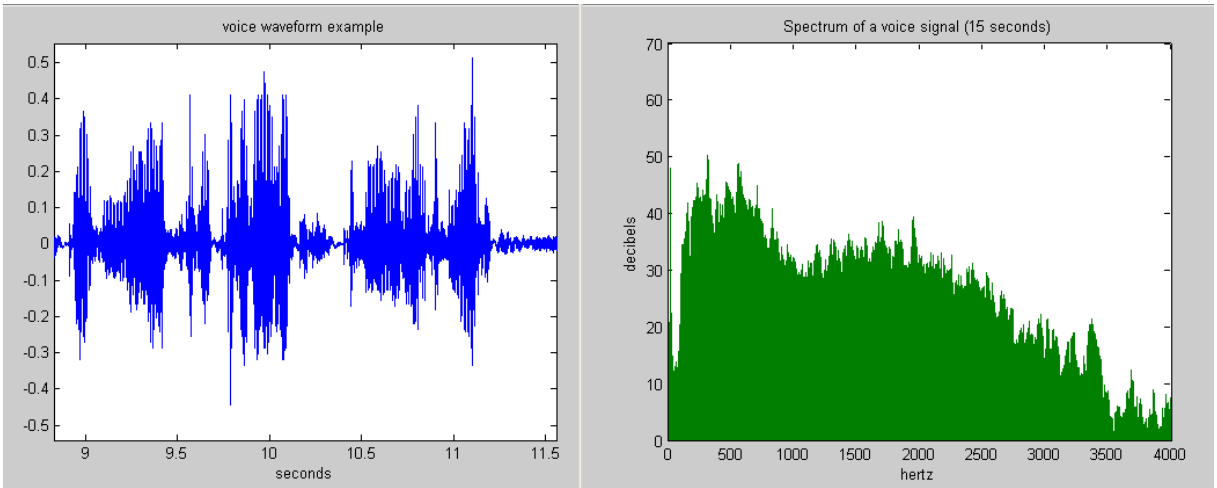


Рисунок А.1 — Пример спектра

б) Частота - количество вхождений повторных событий в единицу времени.

в) Мел — психофизическая единица высоты звука, применяется главным образом в музыкальной акустике. Название происходит от слова «мелодия».
График зависимости высоты звука в мелах от частоты колебаний: **Рисунок А.2**

**Спектрограмма** - это визуальное представление спектра из частот сигнала, изменяющегося в зависимости от времени. Применяется для идентификации речи, анализа звуков животных, в различных областях музыки и т.д.

Общий формат - это график с двумя геометрическими измерениями: одна ось представляет время, а другая частоту; третье измерение,

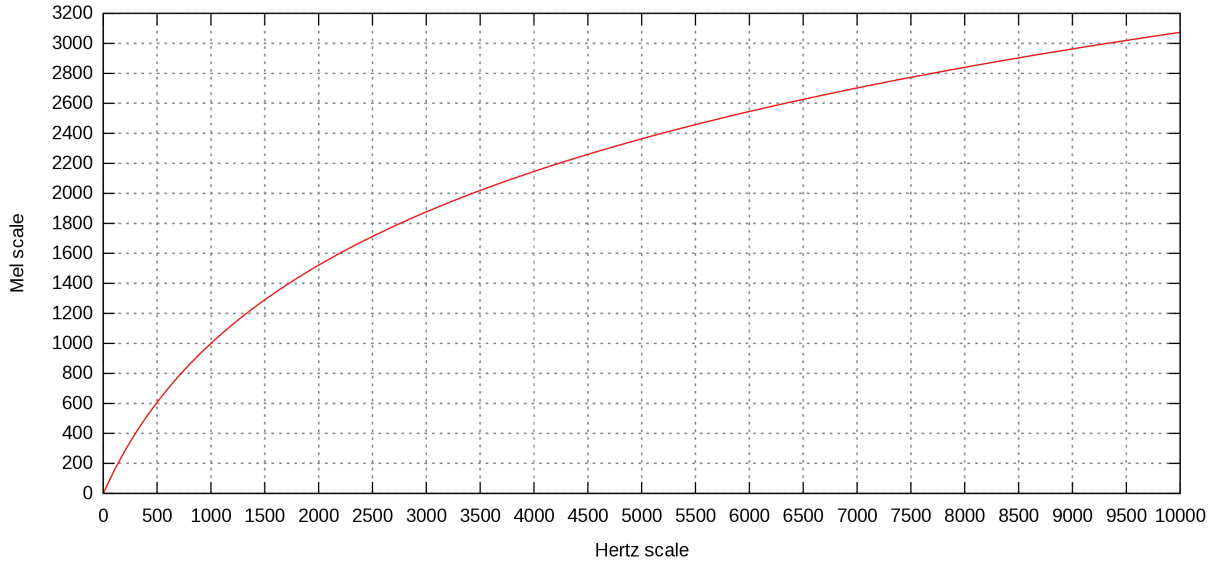


Рисунок А.2 — График зависимости Герц от мелов

показывающее амплитуду конкретной частоты при конкретном t, отличается интенсивностью или цветом каждой точки изображения.

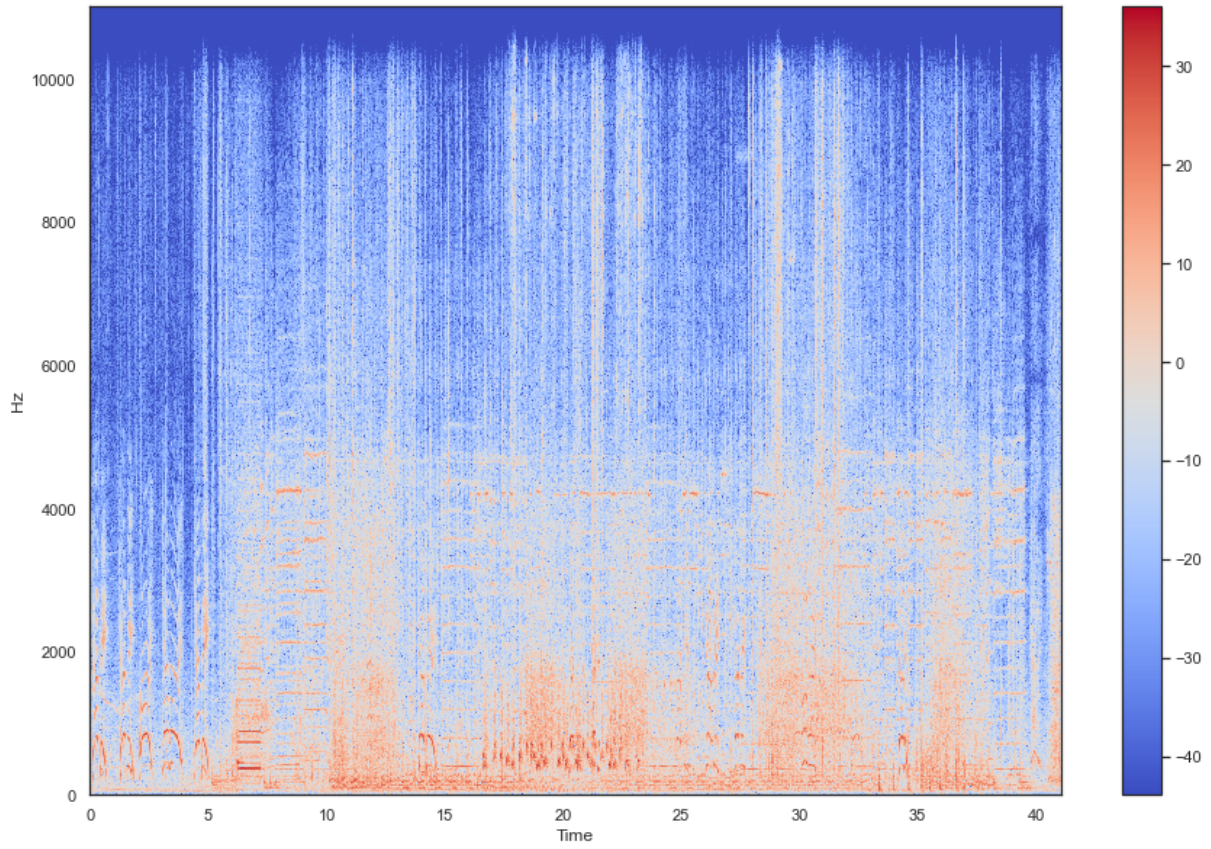


Рисунок А.3 — Спектограмма

**Мел-спектрограмма** — это обычная **спектрограмма**, где частота выражена не в Гц, а в мелах. Переход к мелам осуществляется с помощью применения мел-фильтров к исходной спектрограмме. Мел-фильтры представляют из себя треугольные функции, равномерно распределенные на мел-шкале.

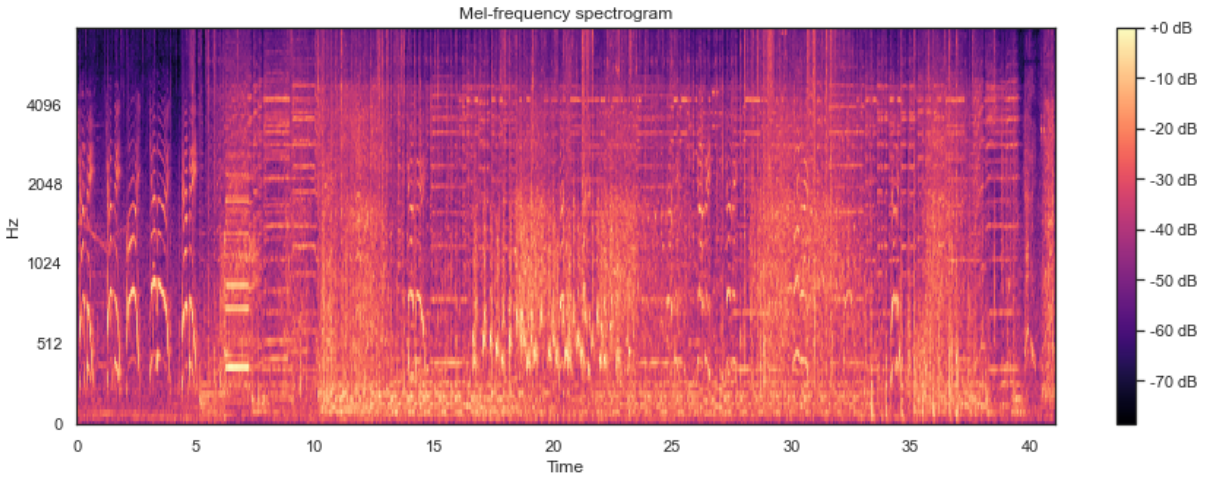


Рисунок А.4 — Спектограмма

**Мел-кепстральные коэффициенты** - представление кратковременного спектра мощности звука, основанное на линейном косинусном преобразовании логарифмического спектра мощности по нелинейной шкале частоты mel.

**OpenSMILE** - интерпретация мультимедиа с открытым исходным кодом с помощью извлечения большого пространства признаков. Также он позволяет извлекать аудио- и видео признаки для обработки сигналов и машинного обучения с акцентом на признаки, позволяющие распознавать эмоции по речи.

**Признаки по умолчанию:**

а) Признаки цветности для распознавания тональностей и аккордов;

б) MFCC(мел-кепстральные коэффициенты) для распознавания речи;

в) PLP для распознавания речи;

PLP очень похож на MFCC. Руководствуясь слуховым восприятием, он использует пре-эмфазис равной громкости и сжатие кубического корня вместо логарифмического сжатия.
Он также использует линейную регрессию для уточнения кепстральных коэффициентов. PLP обладает немного лучшей точностью и лучшей шумостойкостью. Но также считается, что MFCC - это более безопасный выбор.

г) Просодия (высота тона и громкость);

д) Набор признаков INTERSPEECH 2009 Emotion Challenge;

е) Набор признаков паралингвистического вызова INTERSPEECH 2010;

ж) Набор признаков для вызова состояния спикера INTERSPEECH 2011;

з) Набор признаков INTERSPEECH 2012 Speaker Trait Challenge;

и) Набор признаков для сравнения INTERSPEECH 2013;

к) Набор признаков Medieval 2012 TUM для обнаружения сцен насилия;

л) Три эталонных набора признаков для распознавания эмоций (старые наборы, устаревшие из-за новых наборов задач INTERSPEECH);

м) Аудиовизуальные признаки, основанные на аудиофайлах паралингвистического вызова INTERSPEECH 2010.

# Приложение Б Telegram бот

Бота вы можете найти в Telegram под ником @open_MER_bot или перейти по ссылке [https://t.me/open_MER_bot].

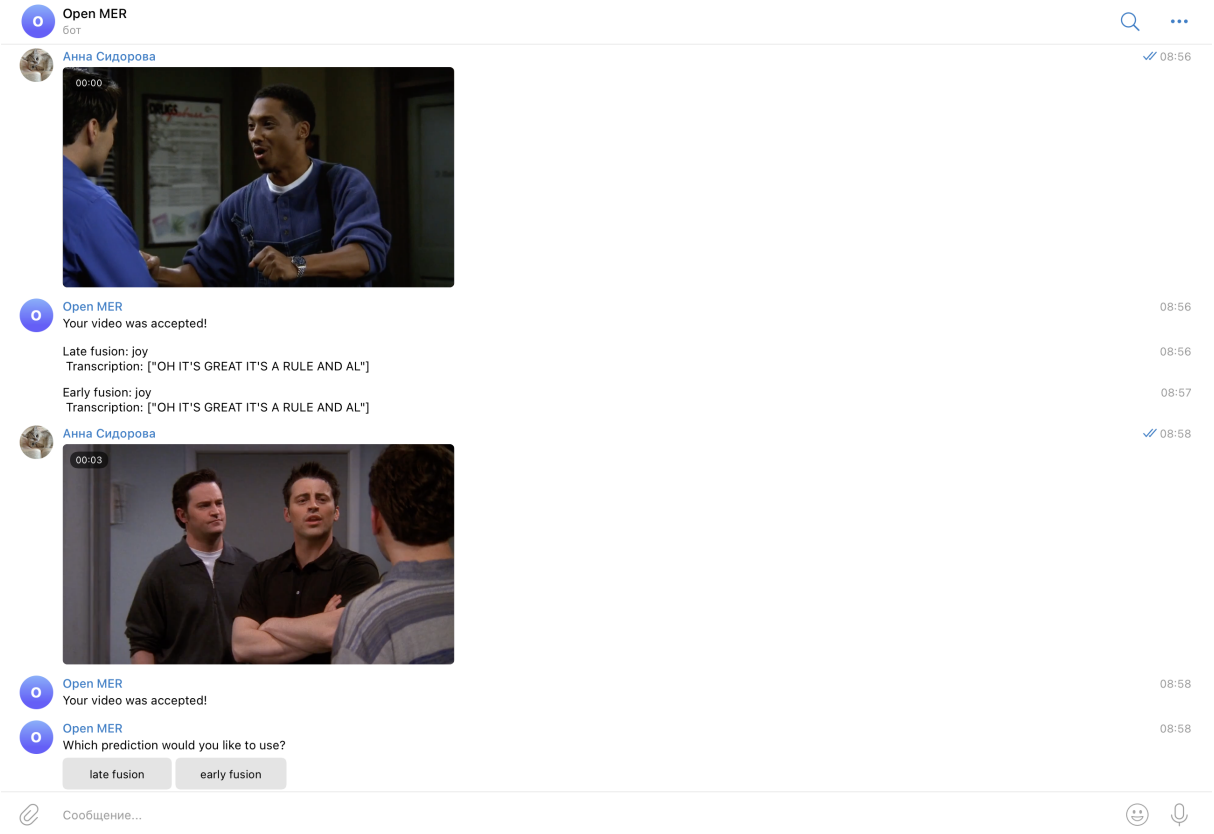


Рисунок Б.1 — Telegram bot