Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное образовательное учреждение высшего образования
«**НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО**»
(Университет ИТМО)

Факультет программной инженерии и компьютерной техники

Пояснительная записка к курсовому проекту

По предмету «**Нейротехнологии и аффективные вычисления**»

Тема проекта: «**Анализ эмоций интервьюируемого человека из разных профессий**»

Выполнили:
Иоган Максим Александрович
P332222
Проверил:
Русак Алена Викторовна

Санкт-Петербург, 2023

Содержание

Команда проекта

Иоган Максим: использовал библиотеки FER, torch, Librosa для анализа эмоций по разным модальностям

Аннотация

Данный проект предполагает исследование взаимосвязи между эмоциями человека и его профессиональной деятельностью на основе мультимодального анализа эмоций. В моей работе я выполняю классификацию по унимодальным типам данных. Чтобы классифицировать одномодальные входные данные, извлеченные одномодальные признаки отправляются в 2 сети предоставленные библиотека FER, PyTorch в качестве входных данных для классификации настроений.
Итоговым результатом данного проекта будет наглядная статистика и визуализация с их анализом, а также вывод о взаимосвязи между эмоциями человека и его профессиональной деятельностью

Введение

На данный момент не имеется научных работ по исследованию настроения человека в ходе интервью среди людей разных сфер деятельности. В результате данной работы мы можем выяснить есть ли зависимость между эмоциональным фоном человека и его работой и в чем различие между людьми из разных областей. На какие вопросы люди реагируют особым образом, а какие воспринимают спокойно.

Область исследования зависимости эмоций человека от его профессии практически не имеет актуальных и обоснованных работ. Все выводы делаются по субъективному ощущению, что не может являться точным способом поиска зависимости. Также мало кто при анализе эмоций основывается сразу на трех факторах: голос, текст, видео, что увеличивает точность предсказания. Большинство исследований нацелены на получение информации об эмоциональной обратной связи человека от какого-либо продукта, действия, предмета, но никто не пытается найти взаимосвязь между эмоциями человека и его жизненным бэкграундом.

В качестве области исследования я буду использовать интервью с канала вДудь с людьми из разных сфер деятельности. Скачать видео и текстовую расшифровку можно без проблем.

Цель

Исследование взаимосвязи между эмоциями и жизненными интересами людей из разных сфер деятельности

Задачи

1. Анализ предметной области
2. Найти тренировочные данные для обучения моделей
3. Обучить модель на тренировочных данных
4. Выгрузить видеоматериалы и текстовые расшифровки интервью с разными людьми для подготовки тестовой выборки
5. Сгруппировать их по общим профессиям или сферам деятельности
6. В рамках каждой группы, подготовить каждое видео (избавиться от фрагментов, которые не несут полезной информации) как тестовую выборку

7. Разделить видео на три компонента: изображение, текст, аудио
8. По каждому из компонентов провести анализ эмоций
9. Объединить предсказание учитывая все три компонента
10. Визуализировать результат на графиках, сводных таблицах
11. После анализа всех материалов, сделать вывод о каждой группе, об их взаимосвязи, о причинах возможных всплесках эмоций.

База данных

CREMA-D — это набор данных из 7442 оригинальных клипов от 91 актера. Эти клипы были сняты 48 мужчинами и 43 женщинами-актёрами в возрасте от 20 до 74 лет, принадлежащими к разным расам и этническим группам. Актеры говорили на выбор из 12 предложений. Предложения были представлены с использованием одной из шести различных эмоций (гнев, отвращение, страх, счастье, нейтральная и грустная) и четырех различных уровней эмоций (низкий, средний, высокий и неопределенный).

RAVDESS содержит 1440 файлов: 60 попыток на актера * 24 актера = 1440. RAVDESS содержит 24 профессиональных актера (12 женщин, 12 мужчин), озвучивающих два лексически совпадающих утверждения с нейтральным североамериканским акцентом. Речевые эмоции включают выражения спокойствия, счастья, грусти, гнева, страха, удивления и отвращения. Каждое выражение производится на двух уровнях эмоциональной интенсивности (нормальный, сильный) с дополнительным нейтральным выражением.

SAVEE была записана от четырех носителей английского языка (обозначенных как DC, JE, JK, KL), аспирантов и исследователей Университета Суррея в возрасте от 27 до 31 года. Эмоции были описаны психологически в дискретных категориях: гнев, отвращение, страх, счастье, печаль и удивление. Также добавлена нейтральная категория для записи 7 категорий эмоций.

TESS Две актрисы (26 и 64 года) произносили набор из 200 целевых слов во фразе-носителе «Скажи слово _ », и были сделаны записи набора, изображающего каждую из семи эмоций (гнев, отвращение, страх, счастье). , приятный сюрприз, грусть и нейтральный) Всего 2800 точек данных (аудиофайлов).

Главная идея

Речевой сигнал сначала преобразуется в различные читаемые физические характеристики (высоту тона, энергию и т.д.) с помощью системы обработки речи. Каждый сегмент речевого сигнала имеет свои уникальные характеристики. Некоторые из этих характеристик будут искусственно выбраны, извлечены системой, разделены на фрагменты в 5 секунд, введены в предварительно обученный классификатор для различения и выведены в результате эмоционального состояния.

Текстовый сигнал отправляется на вход модели RuBERT, которая обучена на 351797 русских текстах размеченных по трем эмоциям - нейтральная, позитивная и негативная. Используя скачанные субтитры с размеченным диапазоном времени, с помощью библиотеки PySrt мы делим текст на 5 секундные интервалы и подаем модели на вход.

Изображение обрабатывается с помощью библиотеки FER, которое определяет выражение лица по фотографии. Каждые 25 кадров берется один кадр и отправляет на обработку библиотеки, которая находит лицо и возвращает значение вероятности принадлежности к одной из семи эмоций (злость, отвращение, страх, счастье, удивление, нейтрально)

Извлечение признаков

Извлечение признаков — очень важная часть анализа и поиска взаимосвязей между разными вещами. Поскольку мы уже знаем, что данные, предоставленные аудио, не могут быть поняты моделями напрямую, поэтому нам необходимо преобразовать их в понятный формат, для которого используется извлечение признаков.

Звуковой сигнал представляет собой трехмерный сигнал, в котором три оси представляют время, амплитуду и частоту. (Рис. 1 Трехмерный сигнал.)

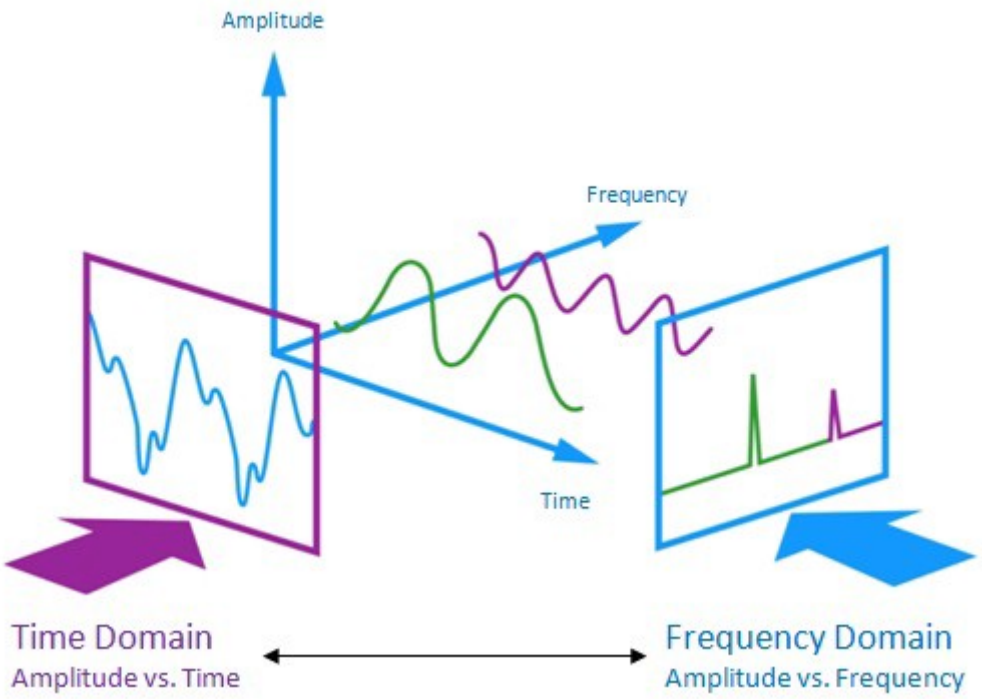

Рис. 1 Трехмерный сигнал

С помощью частоты дискретизации и выборочных данных можно выполнить несколько преобразований, чтобы извлечь из звукового сигнала ценные признаки.

Скорость пересечения нуля: скорость изменения знака сигнала в течение определенного кадра.
Энергия: сумма квадратов значений сигнала, нормализованных по соответствующей длине кадра.
Энтропия энергии: энтропия нормированных энергий подкадров. Его можно интерпретировать как меру резких изменений.
Спектральный центроид: центр тяжести спектра.
Спектральное распространение: Второй центральный момент спектра.
Спектральная энтропия: энтропия нормализованных спектральных энергий для набора подкадров.
Спектральный поток: Квадрат разницы между нормализованными величинами спектров двух последовательных кадров.
Спектральный спад: частота, ниже которой сосредоточено 90% распределения амплитуды спектра.
Кепстральные коэффициенты частоты Mel MFCC: образуют кепстральное представление, в котором полосы частот не являются линейными, а распределяются в соответствии с мел-шкалой.

Вектор цветности: 12-элементное представление спектральной энергии, где ячейки представляют 12 равнотемперированных классов высоты тона музыки западного типа (полутоновый интервал).
Отклонение цветности: стандартное отклонение 12 коэффициентов цветности.

В этом проекте я не углубляюсь в процесс выбора признаков, чтобы проверить, какие признаки подходят для нашего набора данных, а извлекаю только 5 признаков (Рис. 2 Функция извлечения признаков):

- Zero Crossing Rate
- Chroma_stft
- MFCC
- RMS(root mean square) value
- MelSpectogram.

```
def extract_features(data):
    # ZCR
    result = np.array([])
    zcr = np.mean(librosa.feature.zero_crossing_rate(y=data).T, axis=0)
    result=np.hstack((result, zcr)) # stacking horizontally

    # Chroma_stft
    stft = np.abs(librosa.stft(data))
    chroma_stft = np.mean(librosa.feature.chroma_stft(S=stft, sr=sample_rate).T, axis=0)
    result = np.hstack((result, chroma_stft)) # stacking horizontally

    # MFCC
    mfcc = np.mean(librosa.feature.mfcc(y=data, sr=sample_rate).T, axis=0)
    result = np.hstack((result, mfcc)) # stacking horizontally

    # Root Mean Square Value
    rms = np.mean(librosa.feature.rms(y=data).T, axis=0)
    result = np.hstack((result, rms)) # stacking horizontally

    # MelSpectogram
    mel = np.mean(librosa.feature.melspectrogram(y=data, sr=sample_rate).T, axis=0)
    result = np.hstack((result, mel)) # stacking horizontally

    return result
```

Рис. 2 Функция извлечения признаков

Перед отправкой текстовых данных, мы токенизируем их, представляем каждое предложение в виде вектора, где каждое значение это количество слов в данном предложении, а вектор имеет длину равную количество уникальных слов из всего набора предложений.

Извлечением признаков изображения занимается библиотека FER автоматически.

Алгоритмы классификации

Классификация аудио:
Для классификации аудио я решил построить сверточную нейронную сеть (Рис. 3 Топология CNN), но прежде я хотел бы наглядно объяснить как по аудио можно различать эмоции (Рис. 4

Волновые графики голоса). При гневе вся фраза как будто бы произносится на одном дыхании, с большим напором в начале, это отчетливо видно на графике. Если посмотреть на график «счастья», то можно заметить, что фраза произносится протяжнее, равномернее. Видно, что на записях присутствуют небольшие шумы. Именно изменение интонации и тембра голоса поможет нам классифицировать эмоции, представленные на аудиозаписях.

Layer (type)	Output Shape	Param #
conv1d_28 (Conv1D)	(None, 162, 256)	1536
max_pooling1d_28 (MaxPooling	(None, 81, 256)	0
conv1d_29 (Conv1D)	(None, 81, 256)	327936
max_pooling1d_29 (MaxPooling	(None, 41, 256)	0
conv1d_30 (Conv1D)	(None, 41, 128)	163968
max_pooling1d_30 (MaxPooling	(None, 21, 128)	0
dropout_13 (Dropout)	(None, 21, 128)	0
conv1d_31 (Conv1D)	(None, 21, 64)	41024
max_pooling1d_31 (MaxPooling	(None, 11, 64)	0
flatten_7 (Flatten)	(None, 704)	0
dense_13 (Dense)	(None, 32)	22560
dropout_14 (Dropout)	(None, 32)	0
dense_14 (Dense)	(None, 8)	264

Total params: 557,288
Trainable params: 557,288
Non-trainable params: 0

Рис. 3 Топология CNN

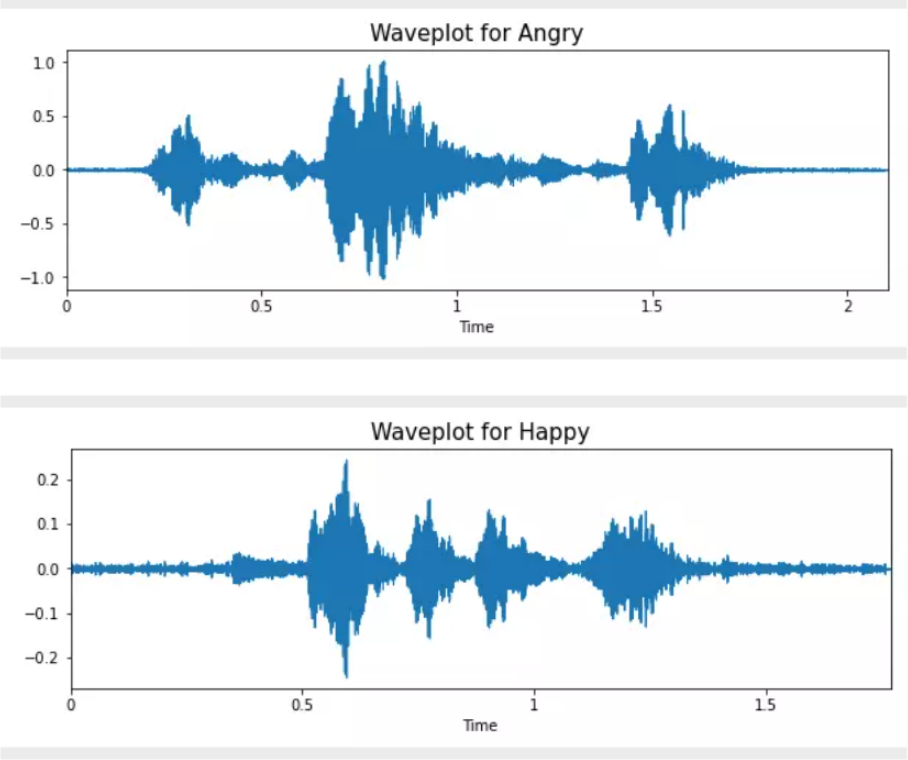

Рис. 4 Волновые графики голоса

Обучать данную сеть я буду на 4-ех датасетах, подробно описанных в разделе [База данных]

- Crowd-sourced Emotional Multimodal Actors Dataset (Crema-D)
- Ryerson Audio-Visual Database of Emotional Speech and Song (Ravdess)
- Surrey Audio-Visual Expressed Emotion (Savee)
- Toronto emotional speech set (Tess)

Графики тренировочной и тестовой метрики accuracy и функции ошибки (Рис. 5 Графики Loss, Accuracy). На контрольной выборке удалось достигнуть точности в 60%.
Посмотрев на отчет по метрикам для каждого класса, можем заметить, что лучше всего классифицируются такие яркие эмоции, как злость и удивление (Рис. 6 Отчет по классификации)

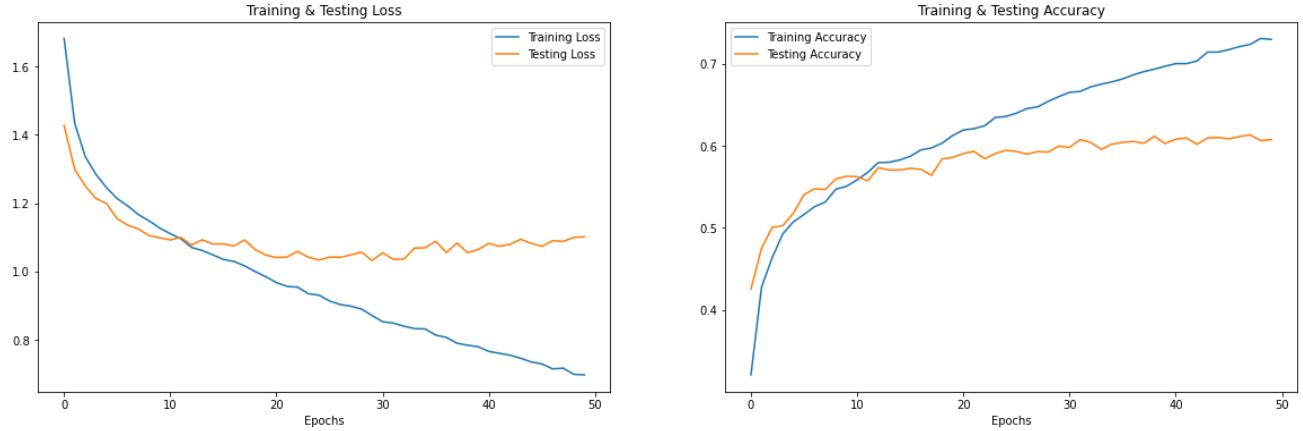

Рис. 5 Графики Loss, Accuracy

	precision	recall	f1-score	support
angry	0.78	0.69	0.73	1396
calm	0.62	0.86	0.72	142
disgust	0.54	0.48	0.51	1461
fear	0.63	0.51	0.57	1443
happy	0.53	0.62	0.57	1450
neutral	0.55	0.57	0.56	1265
sad	0.58	0.68	0.62	1470
surprise	0.85	0.79	0.82	495
accuracy			0.61	9122
macro avg	0.63	0.65	0.64	9122
weighted avg	0.61	0.61	0.61	9122

Рис. 6 Отчет по классификации

Классификация текста:
Для классификации текста я решил использовать обученную на русских текстах модель BERT (это метод машинного обучения на основе преобразователя для предварительной подготовки к обработке естественного языка, разработанный Google), потому что за основу был взят оригинальный код от Google.Так как обучающая выборка содержала 30 GB русского текста, в котором была Википедия, новости, часть корпуса Taiga и немного книг, училась модель около 7—8 дней. Также ruBERT-base имеет BPE-токенизатор со словарём 120 тыс токенов. В модели 178M весов. Соответственно обучать такую модель было бы очень затратно по времени и ресурсам.

Для использовании модели достаточно импортировать ее и токенайзер с помощью библиотеки transformers, указав ссылку на нужную модель. В функции предсказания необходимо предварительно токенизировать текст и передать на вход модели. (Рис. 7 модель ruBERT)

```
tokenizer = BertTokenizerFast.from_pretrained('blanchefort/rubert-base-cased-sentiment')
model = AutoModelForSequenceClassification.from_pretrained('blanchefort/rubert-base-cased-sentiment', return_dict=True)


def predict(text):
    inputs = tokenizer(text, max_length=512, padding=True, truncation=True, return_tensors='pt')
    outputs = model(**inputs)
    predicted = torch.nn.functional.softmax(outputs.logits, dim=1)
    predicted = torch.argmax(predicted, dim=1).numpy()
    return predicted
```

Рис. 7 модель ruBERT

С помощью библиотеки pysrt я работаю с файлом субтитров от видео, что позволяет мне делить текст по длительности времени и подавать на вход модели в цикле (Рис. 7 Процесс классификации текста)

```
decode = {'Neutral':0,'Positive':0,'Negative':0}

subs = pysrt.open('/content/drive/MyDrive/[Russian] Чебатков – стендап для мозга (Eng subs) [DownSub.com].srt')
total = [0,0,0]
i = 0
pbar = tqdm(total = 2148)
while i < 2184:
  j=i+1
  while (subs[j].end.seconds + subs[j].end.minutes * 60 + subs[j].end.hours * 60**2) - \
  (subs[i].start.seconds + subs[i].start.minutes * 60 + subs[i].start.hours * 60**2) < 4 \
  and j < 2184:
    j+=1
  total[predict(subs[i:j].text)[0]] += 1
  i = j
  pbar.update(1)
pbar.close()
```

Рис. 7 Процесс классификации текста

Классификация изображения:
Для классификации изображения я использовал предобученную модель из библиотеки FER, потому что обучать собственную модель слишком затратно по времени и ресурсам и точность может быть намного ниже моделей, которые были построены профессионалами и обучены на большом количестве данных.

```
# Вставить расположение видеофайла, который должен быть обработан
location_videofile = "drive/MyDrive/Чебатков – стендап для мозга (Eng subs).mp4"

# Создайте детектор обнаружения лиц
face_detector = FER(mtcnn=True)
# Введите видео для обработки
input_video = Video(location_videofile)

# Фннкция Analyze() запустит анализ на каждый 100ый кадр входного видео
# Он создаст прямоугольную рамку вокруг каждого изображения и покажет значения эмоций рядом с ней.
# Наконец, метод опубликует новое видео с рамкой вокруг лица человека с текущими значениями эмоций.
processing_data = input_video.analyze(face_detector, display=False, frequency=100)

# Теперь мы преобразуем проанализированную информацию в фрейм данных.
# Это поможет нам импортировать данные в виде файла .CSV для последующего анализа.
vid_df = input_video.to_pandas(processing_data)
vid_df = input_video.get_first_face(vid_df)
vid_df = input_video.get_emotions(vid_df)

# График эмоций в зависимости от времени в видео
pltfig = vid_df.plot(figsize=(20, 8), fontsize=16).get_figure()

# Теперь мы будем работать с кадром данных, чтобы определить, какая эмоция была заметна в видео.
angry = sum(vid_df.angry)
disgust = sum(vid_df.disgust)
fear = sum(vid_df.fear)
happy = sum(vid_df.happy)
sad = sum(vid_df.sad)
surprise = sum(vid_df.surprise)
neutral = sum(vid_df.neutral)

emotions = ['Angry', 'Disgust', 'Fear', 'Happy', 'Sad', 'Surprise', 'Neutral']
emotions_values = [angry, disgust, fear, happy, sad, surprise, neutral]

score_comparisons = pd.DataFrame(emotions, columns = ['Human Emotions'])
score_comparisons['Emotion Value from the Video'] = emotions_values
score_comparisons
```

Рис. 8 Классификация изображений

Результаты проекта

Рассмотрим результаты анализа для юмористов.

Анализ эмоций по изображению показал, что у людей из сферы юмора более всего преобладает нейтральное и счастливое выражение лица (Рис. 9 График эмоций комиков по изображению)

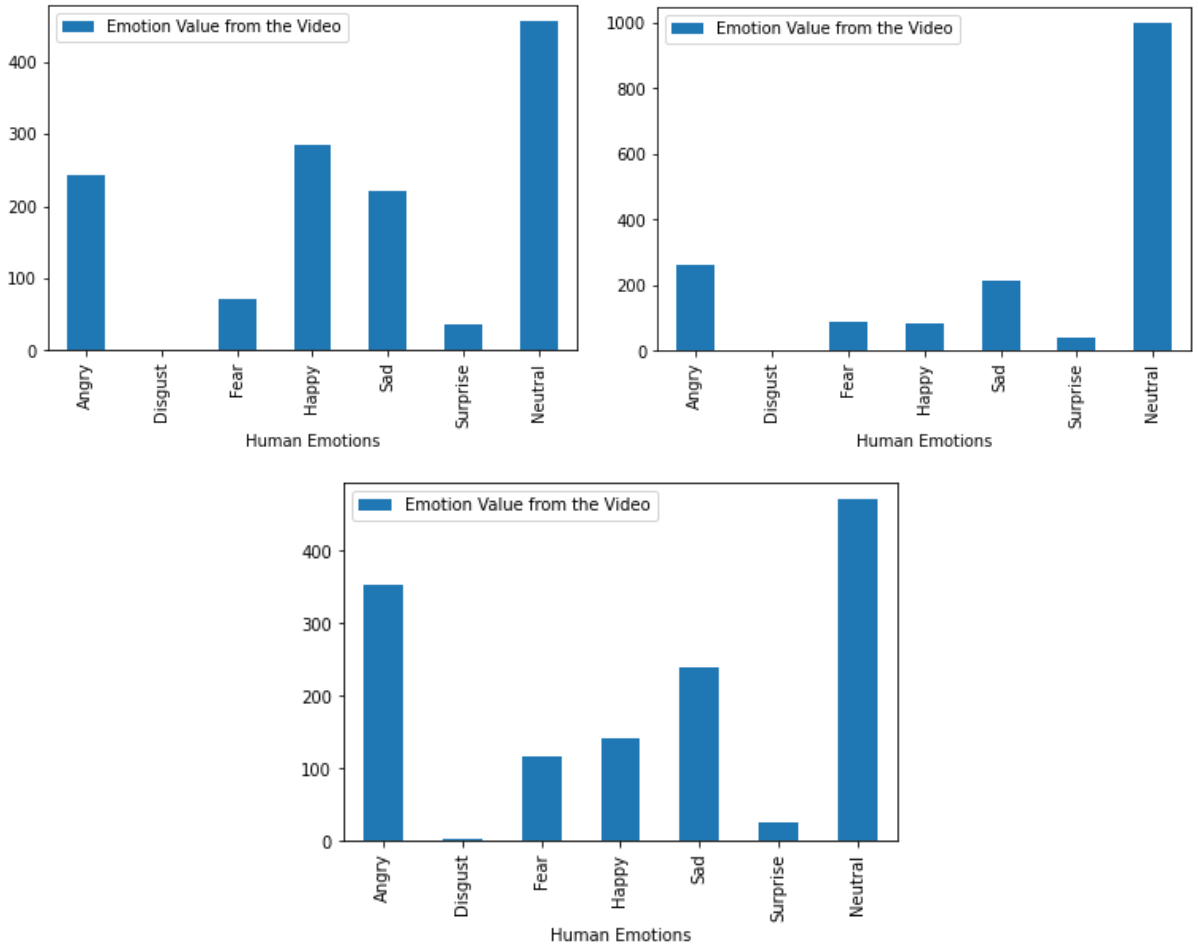

Рис. 9 График эмоций комиков, певцов, бизнесменов по изображению (соответственно)

Анализ эмоций по аудио показал, что эмоция агрессии в голосе преобладает над всеми остальными (Рис. 10 График эмоций комиков по аудио)

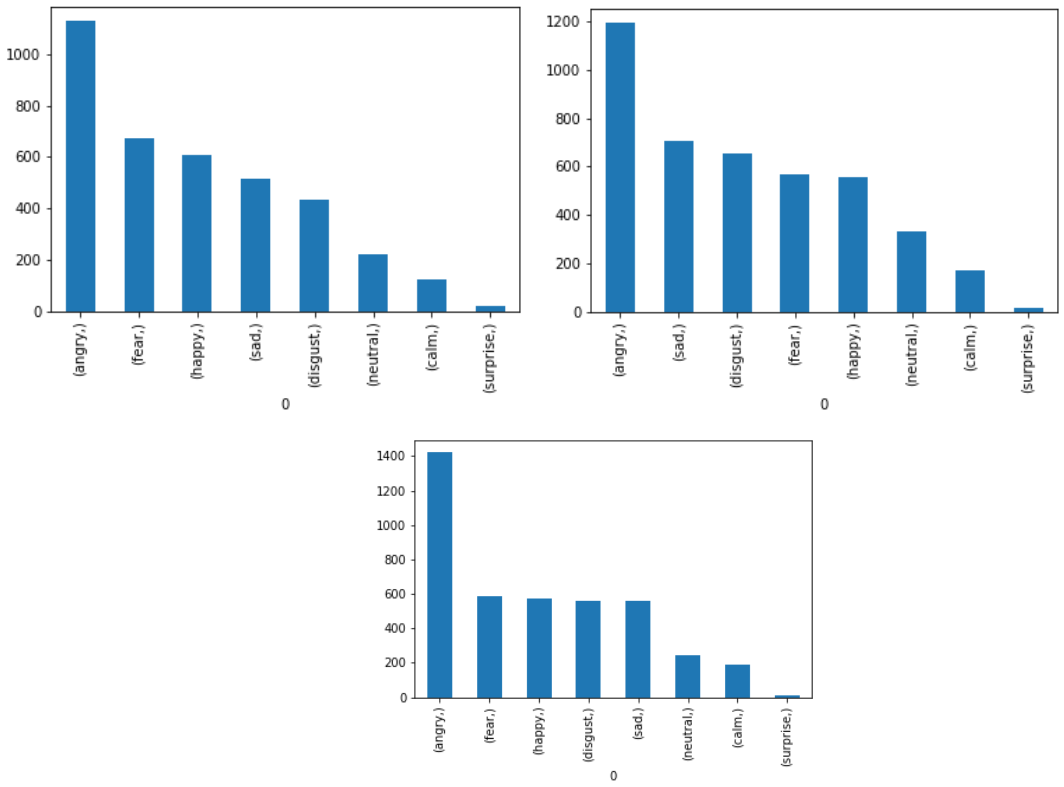

Рис. 10 График эмоций комиков, певцов, бизнесменов по аудио (соответственно)

Анализ эмоций по тексту показал, что эмоция агрессии в тексте преобладает над всеми остальными (Рис. 11 График эмоций комиков по тексту)

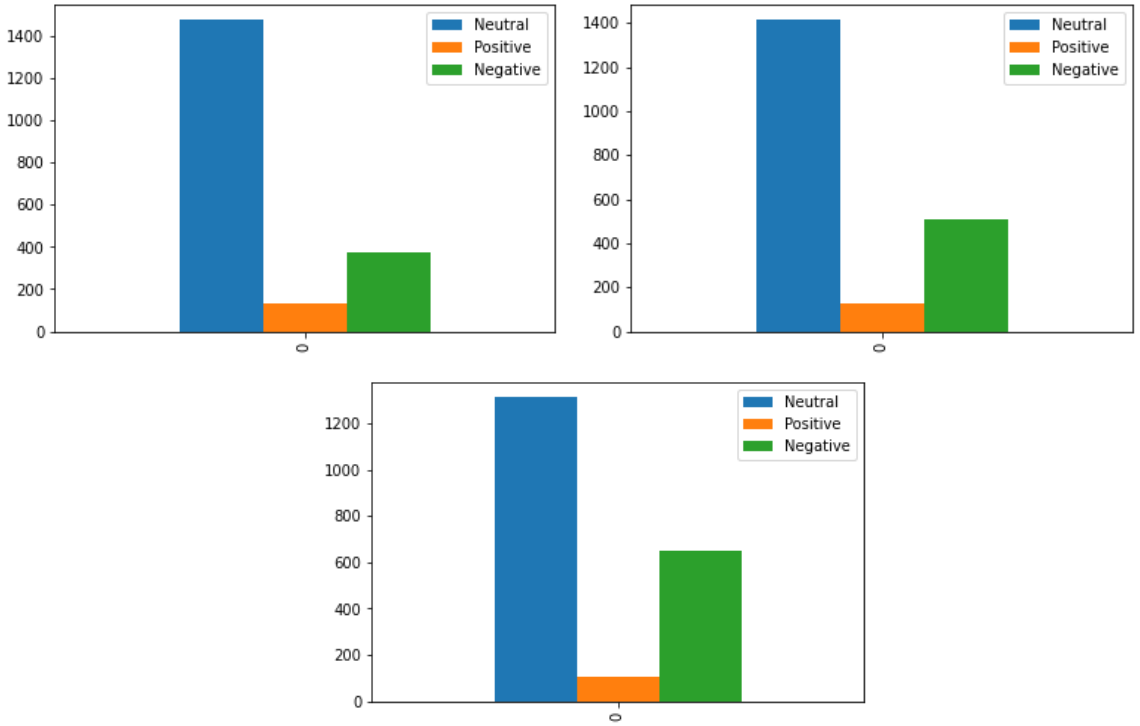

Рис. 11 График эмоций комиков, певцов, бизнесменов по тексту (соответственно)

Вывод

В процессе реализации данного проекта я написал уникальный скрипт, который принимает на вход видео и выдает информацию о том, в течении какого времени человек находился в том или ином эмоциональном состоянии опирая на три модальности: голос, текст, изображение. С помощью этого мы определили чем отличаются на эмоциональном уровне люди из разных профессий и отличаются ли вообще.

Для начала разберем результаты анализа по изображению. Заметим, что у всех групп преобладает нейтральное состояние, что достаточно правдоподобно, так как человек не может долго поддерживать сильные эмоции, такие как радость или гнев и большую часть времени испытывает нейтральные эмоции. Поэтому мы обратим внимание на следующую по количеству эмоцию. Для комиков - это счастье, что может означать, что люди из юмористических сфер деятельность склонны больше улыбаться, шутить. Также важно отметить, что эмоция агрессии очень близко по уровню с счастьем. Это можно интерпретировать, как качельное эмоциональное состояния, человек переходит из злобы к счастью и на контрастах этих эмоций оказывает эффект на свою публику.

Следующий график показывает нам, что певцы в большой степени пребывают в нейтральном состоянии, а все остальные эмоции находятся на минимуме. Это интересное замечание, учитывая, что и юмористы и певцы работают на сцене перед публикой, однако эмоции, которые певцы испытывают в жизни, скуднее, чем у комиков.

График бизнесменов показывает большое значение в столбце агрессии. Если обратиться к тому образу жизни, которым приходится жить бизнесменам, то можно заметить, что им в большинстве своем нужно “идти по головам”, быть настойчивыми, упорными, добиваться своего и все это будет отражаться, как злость на их лице. И несмотря на то, что они живут в достатке, это не приносит им большого счастья и это было обосновано ученым тем, что максимальное счастье человек получает от того момента, когда материальные доходы позволяют закрыть базовые нужды, а все что приобретается от изобилия не приносит практически никакого счастья.

Теперь посмотрим на результаты анализа аудио. Во всех случаях преобладает агрессия, причиной чего может быть волнение гостей, быстрота, громкость их речи, что воспринималось алгоритмом, как агрессивные звуки. Также хотел бы заметить, что обученный алгоритм не обладал высокой точностью и результат может быть искажен, поэтому очень сложно сделать какой-либо вывод по полученным графикам.

Посмотрев на результаты анализа текста, можно заметить схожесть с результатами анализа изображений, так как здесь тоже преобладает нейтральное состояние. Позитивный оттенок предложений примерно у всех трех групп одинаковый, однако негативный преобладает у бизнесменов. На втором месте певцы и меньше всего негатива у комиков.

В заключении можно сказать, что по результатам исследования каждая группа имела свои эмоциональные особенности, которые тем или иным образом объяснялись их профессиональной деятельностью. Однако стоит учесть, что нельзя делать выводы по полученным результатам, так как в анализе было задействовано не достаточное количество примеров, в видео могли находиться рекламные вставки, лицо ведущего, разные ракурсы. Все это ухудшает качество предсказания, но то, как эмоции соответствовали профессиональной деятельности интервьюируемых, доказывает, что качество исследования выше случайного предсказания.

Список литературы

1. Распознавание эмоций с использованием нейронной сети : сайт. – URL: https://vc.ru/dev/239027-raspoznavanie-emociy-s-ispolzovaniem-neyronnoy-seti (дата обращения: 04.01.2023)
2. The Ultimate Guide to Emotion Recognition from Facial Expressions using Python : сайт. – URL: https://towardsdatascience.com/the-ultimate-guide-to-emotion-recognition-from-facial-expressions-using-python-64e58d4324ff (дата обращения: 04.01.2023)
3. RuBERT for Sentiment Analysis : сайт. – URL: https://huggingface.co/blanchefort/rubert-base-cased-sentiment (дата обращения: 04.01.2023)
4. Введение в библиотеку Transformers и платформу Hugging Face : сайт. – URL: https://habr.com/ru/post/704592/ (дата обращения: 04.01.2023)