I'm Beside You

Report

Topic:

Introduction videos Evaluation

Prepared by: Adarsh Santoria
Institution: IIT Mandi

Index

Introduction

Project Background

In this report, we analyze a dataset containing emotion scores, transcript scores, and transcripts extracted from the introduction videos of 10 candidates. Our goal is to generate valuable insights to help make informed decisions about candidate recruitment. We will evaluate the candidates' communication skills, identify areas of expertise, and determine whether they are suitable for recruitment based on the provided data.

This project aims to enhance the candidate selection process by leveraging data analysis techniques to identify suitable candidates effectively.

Objectives

Suitable Candidate Identification: Employ data-driven analysis, focusing on emotions and transcripts, to pinpoint candidates aligning closely with role requirements.

Informed Decision-Making: Provide actionable insights derived from candidate data to empower informed decision-making during recruitment.

Data Description

Emotion Data

movie_id: Unique ID for each video.
image_seq: Image number.
angry: Emotion score for anger.
disgust: Emotion score for disgust.
fear: Emotion score for fear.
happy: Emotion score for happiness.
sad: Emotion score for sadness.
surprise: Emotion score for a surprise.
neutral: Emotion score for neutrality.
dominant_emotion: Dominant emotion among all emotions.

gaze: Indicator (1 for looking at the camera, 0 for not).
blink: Eye blink status (1 for blink, 0 for no blink).
eye_offset: Deviation of the eye from the camera.
elapsed_time: Timestamp in seconds (e.g., 9 seconds).

Transcript Data

start: Start time of the spoken text.
end: End time of the spoken text.
text: Transcribed text spoken within the duration.
positive: Positive sentiment score.
negative: Negative sentiment score.
neutral: Neutral sentiment score.
confident: Confidence score.
hesitant: Hesitance score.
concise: Conciseness score.

Exploratory Data Analysis (EDA)

Text Mining

With the help of prompt engineering, skillset data, education data, experience data,etc are are generated as shown below -

Participant 1:

Education: Pursuing postgraduate and management from IIM Kozhikode, B.Tech in Biotechnology from Heritage Institute of Technology Kolkata, M.Tech from IIT Kharagpur.
Work Experience: Three years in regulatory affairs in the pharmaceutical industry, medical writer at Ciro Klein Farm, specialized in drug safety and risk management.
Skills: Regulatory affairs, medical writing, drug safety, risk management, biotechnology.

Participant 2:

Education: Completed BBA in 2022.
Work Experience: Interned in a boutique investment bank, interned with Kabadi Techno, focused on finance.
Skills: Investment banking, finance, financial modeling, startup growth.

Participant 3:

Education: School activity head, top 1.2% rank in undergraduate entrance exam, honours degree from Varanasi University.
Work Experience: Internship in a steel manufacturing firm, internship in an accounting firm.
Skills: Sales, accounting, event coordination.

Participant 4:

Education: Engineering graduate in electronics and communication.
Work Experience: Academic advisor in a school for 19 months.
Skills: Academic advising, electronics, communication.

Participant 5:

Education: Undergraduate in mass media with a specialization in advertising.
Certifications: Entrepreneurship course by Turnip, foundations of management by Google.
Skills: Advertising, entrepreneurship, content creation (drawing, painting, singing, writing movie reviews).

Participant 6:

Education: First-year MBA analytics student from IIM Kashipur, engineering graduate.
Work Experience: Three years at Deloitte, validation processes for pharmaceutical software.
Skills: Analytics, strategy, validation processes.

Participant 7:

Education: Undergraduation in earth science, worked at General Insurance Corporation of India.
Skills: Reinsurance, underwriting, verbal and written communication.

Participant 8:

Education: PGP finance student at IIM Co-Ecode, chartered accountant, CFA Level 1.
Work Experience: Over three years with PwC, internship with ITC Limited. Skills: Financial analysis, internal audit, edtech.

Participant 9:

Education: First-year MBA student at IIM Lucknow, B.Tech in Agriculture Engineering, M.Tech in Food Process Engineering.
Entrepreneurship: Co-founded an Agritech startup.
Project Leadership: Led a project on the application of remote sensing IoT and artificial intelligence in agriculture.
Skills: Agriculture, entrepreneurship, business development, strategy, remote sensing, IoT, AI.

Participant 10:

Education: Graduation in B.Com Honours. Work Experience: Interned as an Accounting Associate and Tax Associate.
Leadership: Captain of the students committee in school, class representative in college, part of Bad Scouts and Guide.
Skills: Accounting, taxation, leadership, social activities.

This step doesn't give any good parameter of visualization as all participants are having different types of skills or qualifications. Hence we are unable to differentiate in between them.

Data Preprocessing

Loading Data:

Data is loaded from CSV files located in 'transcript_data' and 'emotion_data' folders. The 'neutral' column conflict is resolved by renaming columns ('neutral_x' and 'neutral_y'). Ensuring participant_id uniqueness in emotion_data by taking only the first participant_id in emotion_data of a participant to void conflict

Merging Data:

For each row in the transcript_data, filter emotion_data based on elapsed_time such that it will lie in between start and end of transcript_data. This is to avoid unnecessary copies. Then calculate mean values for numeric columns in emotion_data of that range and add it to that row of transcript data . If data is non numeric, take the first column value. All textual data are constant for emotion data except dominant emotion. So the dominant equation is calculated by comparing values of emotion of each row and changed accordingly. Append the resulting data as new rows to 'unfiltered_data' DataFrame. Participant_id is then set as the index of the dataframe (candidate key).

Data Reduction:

- Dropping columns that are always constant (‘temperature’ and ‘distance’).
- Columns like 'avg_logprob,' 'compression_ratio,' 'seek,' 'upload_time,' 'id,' 'image_seq,' and 'movie_id' are dropped as they are not relevant to the analysis or contain little variation.
- Dropping faulty columns like ‘gaze’ and ‘blink’ as it looks too unrealistic while going through data.

Text Analysis

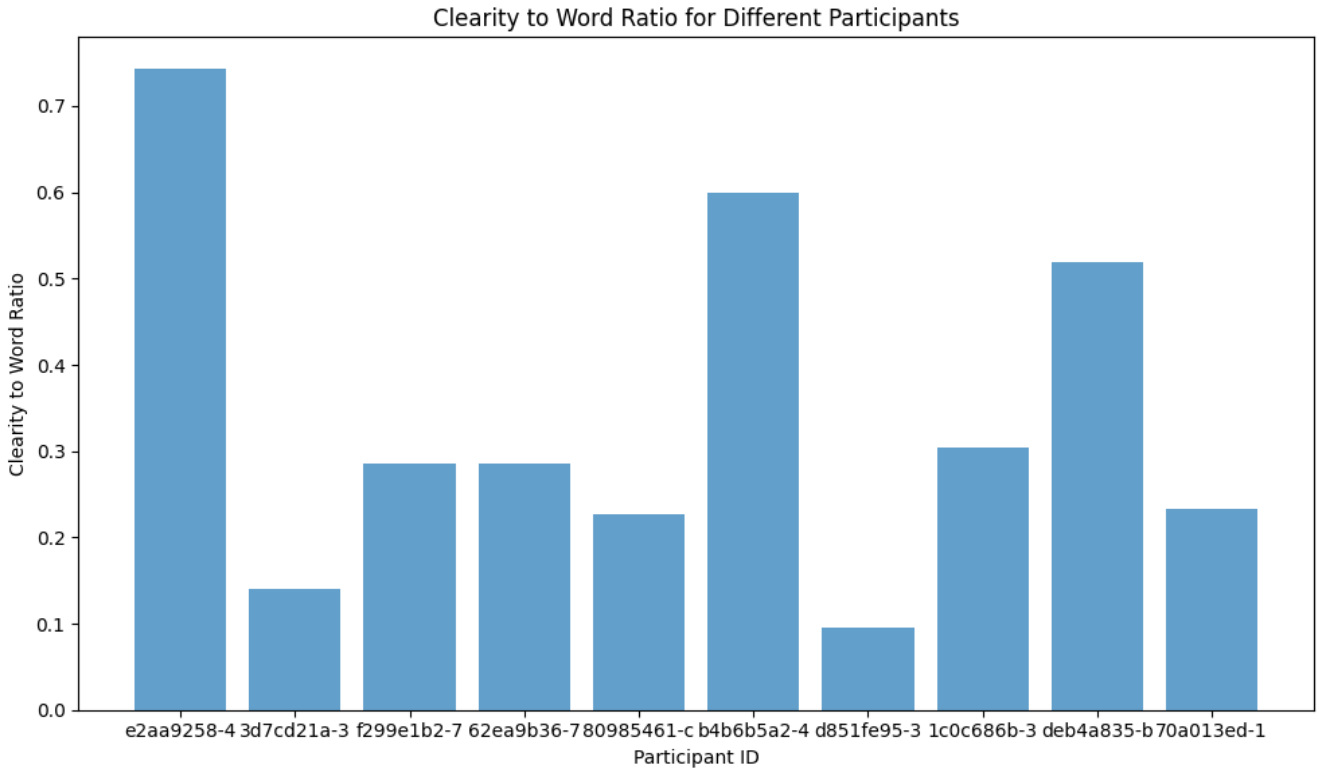
Clearity:

It defines how clearer the words are. By multiplying the time duration (start-end) with speech probability (which is 1-no_speech_prob), we get the time duration of speech. If we multiply it with speech speed (words per second), we will get an

experimental number of words that is not always equal to the actual number of words. Clearity defines the difference between these by the actual number of words.

```
Clearity = time_duration * probability of speech - words_spoken
```

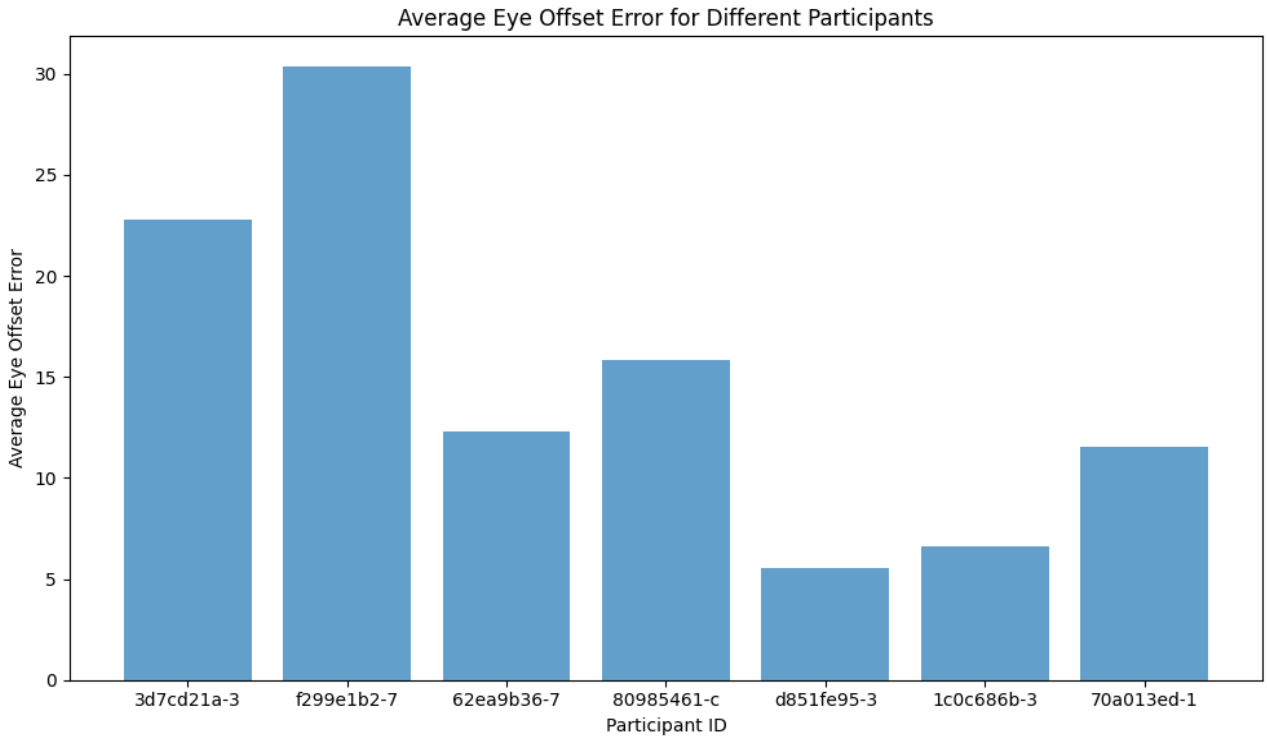
More will be the clearity, less clear words will be. If we take the summation of absolute clearity by sum of the total number of words, we get the knowledge how clear the speech is. If it is not clear, no matter what the participant is telling.



As a threshold 0.5 (which is a lot), if clearity is greater than it we can drop these participants. Participants with participant_id 'E2aa9258-47a5-46ab-9c5c-283460f7a807' , 'b4b6b5a2-4203-41c2-b703-c424dae1fe2b' and 'deb4a835-b82f-4f3d-b2c4-77c66eca7752' are dropped.

Eye Offset:

Average eye _offset of each participant is calculated and if it is greater than 30 as a threshold, participants will be dropped. As this means the participant is looking somewhere else in the time.

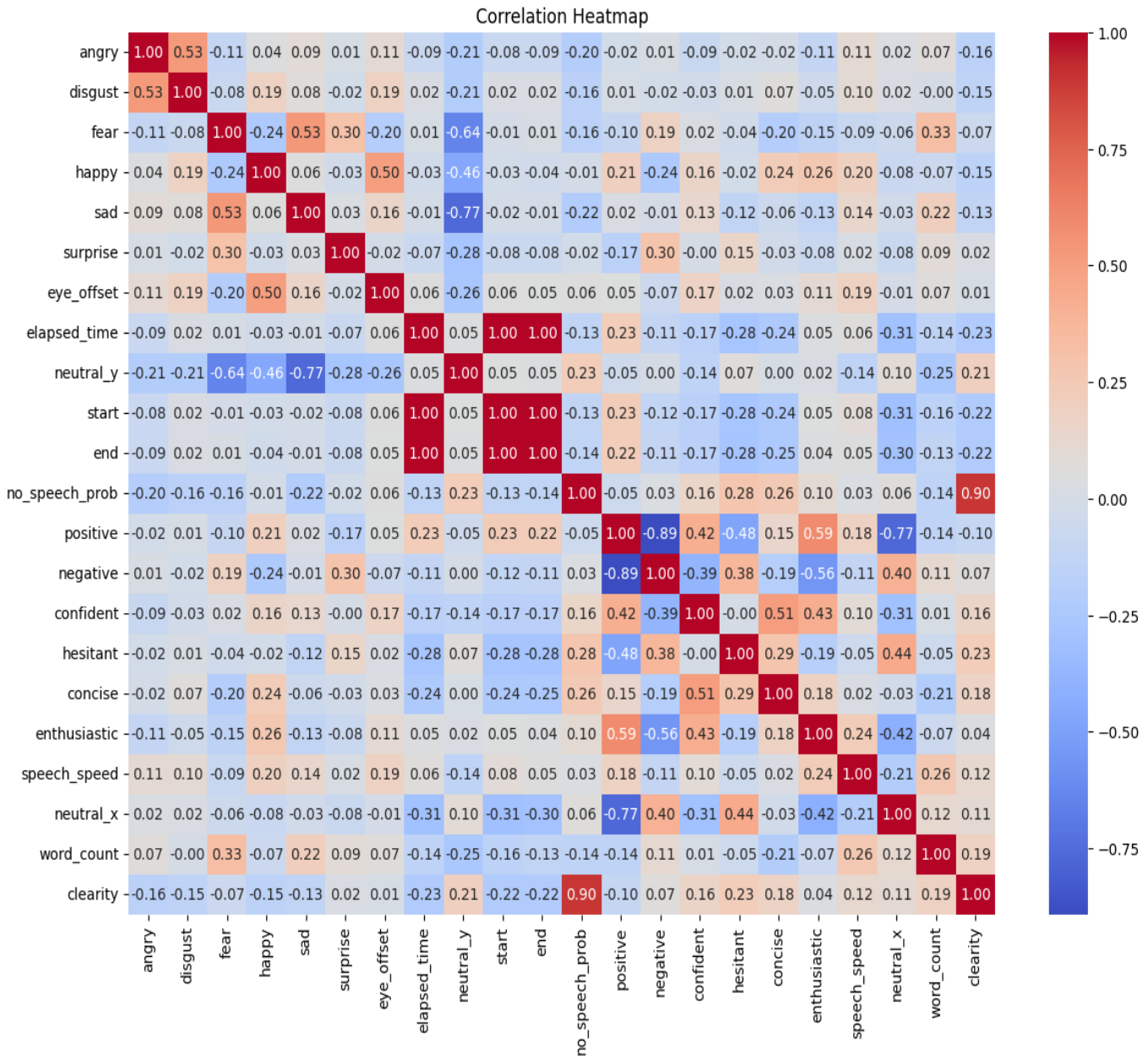


Considering this participant with participant_id 'f299e1b2-7d92-4420-9c5a-d0d2590abdbe' is dropped.

Correlation Analysis

With the help of Correlation Heat Map we can infer the correlation of each attribute with another. Taking the sum of rows of the matrix will give the exact value of the measure. Lesser the value means more correlation. Taking the threshold value 1.5 (analyzing the data), if the threshold is between -1.1 and 1.1, that column will be dropped.

Using this, the 'disgust' column is dropped.



Dominant Data Analysis of Emotion Data

The Dominant Data contains the dominant emotion shown by the participant with emotional scores. For every participant, emotion scores are calculated and emotion with max value is taken as the dominant emotion of the overall pitch of the participant.

Result is displayed as shown below: -

	participant_id	dominant_emotion
0	3d7cd21a-3170-4352-b499-24ea04eaf48c3d7cd21a-3...	happy
1	62ea9b36-7860-4dc9-827c-60060428657162ea9b36-7...	neutral_y
2	80985461-c5d6-466f-a30a-4de2784ed0a380985461-c...	neutral_y
3	d851fe95-3ead-47c1-88aa-d6fc453f7021d851fe95-3...	fear
4	1c0c686b-3aae-4ac6-8625-3e86a7a0892f1c0c686b-3...	neutral_y
5	70a013ed-120a-41fa-bedd-75a5d15afb7670a013ed-1...	fear

Text Analysis and Feature Engineering

Sentiment Analysis

Like Emotion and Transcript Data, we have calculated sentiment_score from Transcript text using the vader sentiment analysis library. This helps in providing more insights from given data. We have also created an avg_eye_offset column because while explaining eyes may move but average must be nearer to 0 (towards camera). Similarly we have taken average of other emotion characteristics. These are compound scores as it contains the overall sentiment of the speech.

Customized Scoring System

We calculated the normalised_eye_offset by min-max normalization of avg_eye_offset and introduced a new column emotional_attitude_score which stores the mean of all the emotional characteristics scores. Similarly we have taken a clearity score which is equal to 1-clearity/word_count as clearity column is representing the number of extra words spoken or not spoken. Hence 1- average error is taken as a clearity score as it determines the clearity in speech.

Hyperparameter Tuning

Grid Search for Hyperparameters

To create a comprehensive evaluation system, we introduced a custom transformer class for weighted scoring and a parameter grid. This step allowed us to assign different weights to various attributes, including sentiment, emotional_attitude_score, clearity_score, and normalised_eye_offset which then gives the best weights parameter grid. These weighted scores with best weights would provide a holistic view of each candidate's introduction as an output final_score which is the mean of the row formed by applying Grid Search on a grid of those four parameters.

Candidate Ranking and Final Scoring

The above graph is made after sorting the participant ids for getting better visualization upto which participant we have to choose from starting.

	participant_id	mean_final_score
0	1c0c686b-3aae-4ac6-8625-3e86a7a0892f	0.629806
1	3d7cd21a-3170-4352-b499-24ea04eaf48c	0.312557
2	62ea9b36-7860-4dc9-827c-600604286571	0.481754
3	70a013ed-120a-41fa-bedd-75a5d15afb76	0.551848
4	80985461-c5d6-466f-a30a-4de2784ed0a3	1.158324
5	d851fe95-3ead-47c1-88aa-d6fc453f7021	0.829686

For the threshold, the best suitable value looks 0.8. This allows participants with participant id '80985461-c5d6-466f-a30a-4de2784ed0a3' and 'd851fe95-3ead-47c1-88aa-d6fc453f7021' are chosen as a result.

Conclusion and Insights

In conclusion, this report has outlined a comprehensive approach to evaluating candidates based on their introduction videos like considering clearity of words and removing high avg_offset participants. By combining sentiment analysis, emotional attitude assessment, clarity evaluation, and eye contact measurement, a customized scoring system was created to provide a holistic view of each candidate's communication skills.

Through hyperparameter tuning and grid search, the best weights for each attribute were determined, resulting in a final score that represents the overall quality of the candidate's introduction. This approach enables informed decision-making during the recruitment process.

Based on the final scores and an approximate threshold, candidates with participant IDs '80985461-c5d6-466f-a30a-4de2784ed0a3' and 'd851fe95-3ead-47c1-88aa-d6fc453f7021'**(Sakshi and Joseph Nichols)** were identified as strong candidates for further consideration. This data-driven analysis

enhances the candidate selection process, ultimately leading to more informed and effective recruitment decisions.