Exploratory Data Analysis Report

Evaluation of Introduction Videos

Spriha Ghosh

28.09.2023

Table of Contents

1. Objective of Exploratory Data Analysis 4

2. Data Overview 4

3. Data Cleaning 5

4. Text Analysis 7

5. Data Visualizations and Insights 8

6. Final Insights 25

Objective of Exploratory Data Analysis

In the context of our analysis, we have been provided with a dataset comprising information from 10 candidates. This dataset encompasses emotion scores, transcript scores, and the associated transcripts derived from their introductory videos.

Our primary objective is to conduct an in-depth exploratory data analysis (EDA) to extract valuable insights and patterns from this dataset. Through this process, we aim to determine whether the candidates should be recruited or not based on data-driven insights.

We will focus on the comprehensive understanding of the candidates' emotional expressions, analysis of their communication skills, and check for desirable areas of expertise from their education, skills and experience as mentioned in the transcripts.

Data Overview

For each of the candidates, the following data is Provided:

Emotion Score: This dataset contains emotions of the candidate throughout the video. It consists of 3 separate files, namely emotion.csv, gaze.csv, metadata.csv.

1) Emotion.csv –

Column Name	Description	Data Type
movie_id	Unique ID for a video	object
image_seq	Image number	int64
angry	angry emotion score	float64
disgust	disgust emotion score	float64
fear	fear emotion score	float64
happy	happy emotion score	float64
sad	sad emotion score	float64
surprise	surprise emotion score	float64
neutral	neutral emotion score	float64
Dominant_emotion	dominant emotion among other emotions	object

2) Gaze.csv –

Column Name	Description	Data Type
movie_id	Unique ID for a video	object
image_seq	Image number	int64
gaze	Candidate is looking at the camera or not. 1 for looking and 0 for not looking	int64
blink	Eye blink (1 -> blink, 0 -> No blink)	int64
eye_offset	Deviation of eye from the camera	float64

3) Metadata.csv –

Column Name	Description	Data Type
movie_id	unique id for a video	object
image_seq	image number	int64
participant_id		object
elapsed_time	timestamp in seconds (9 means at 9th sec)	float64
upload_time		object
distance		float64

Transcript Score: This dataset contains scores extracted from the transcripts throughout the video.

Column Name	Description	Data Type
id		int64
seek		int64
start	start time of the text	float64
end	end time of the text	float64
text	transcript spoken from start to end duration	object
tokens		object
temperature		float64
avg_logprob		float64
compression_ratio		float64
no_speech_prob		float64
positive	positive score	float64
negative	negative score	float64
neutral	neutral score	float64
confident	confidence score	float64
hesitant	hesitance score	float64
concise	concise score	float64
enthusiastic	enthusiasm score	float64
speech_speed	speed of the speech spoken	float64

Transcript Text: Text files containing the actual text of the transcript of the each of the candidate's video is given

Data Cleaning

(Codes for Data Cleaning)

1) Removal of unnecessary columns –

The columns removed from the metadata.csv file in the emotion_data folder is: *participant_id*, *upload_time* and *distance*.
The columns removed from the files in the transcript_data folder is: *seek, tokens, temperature, avg_logprob, compression_ratio* and *no_speech_prob*.

These columns will not be necessary for the analysing the emotion and transcript scores.

The modified tables look like the following –
metadata.csv

image_seq	elapsed_time
6	7
7	8
11	12

Transcript_data table

id	start	end	text	positive	negative	neutral	confident	hesitant	concise	enthusiast	speech_speed
0	0	5.56	Hello, I an	0.580265	0.152281	0.267454	0.846701	0.845698	0.635805	0.647783	2.517986
1	5.56	9.6	IIM Coiko	0.550327	0.189263	0.26041	0.679283	0.733701	0.544145	0.41739	3.217822
2	9.6	14.48	Technolog	0.63986	0.11115	0.24899	0.902729	0.83462	0.715861	0.700062	2.868852

2) Finding missing values –

All the tables (emotion.csv, gaze.csv, metadata.csv, transcript_data tables) are traversed and **no missing value** in any row is found.

3) Finding and Removal of duplicate rows –

All the tables (emotion.csv, gaze.csv, metadata.csv, transcript_data tables) are traversed and **no duplicate rows** are found.

4) Combining separate emotion score tables –

As for every candidate, the emotion score is spread throughout three different files (emotion.csv, gaze.csv, metadata.csv), we will create merged_emotion_1, merged_emotion_2 etc. for all the candidates by merging the mentioned files. "image_seq" column is used as the common field for the merging process. The duplicate column (movie_id) is displayed only once.

Merged_emotion_i table

movie_id	image_seq	angry	disgust	fear	happy	sad	surprise	neutral	dominant_emotion	gaze	blink	eye_offset	elapsed_time
93663f94-	6	6.41279	0.000239	4.53791	0.134349	3.56569	0.555717	84.7933	neutral	0	0	26.8643	7
93663f94-	7	29.8132	1.36594	31.5051	5.55513	11.357	2.18964	18.214	fear	1	0	1.9027	8
93663f94-	11	37.8121	0.001945	2.65793	4.88018	5.27883	4.57332	44.7957	neutral	0	0	32.5426	12
93663f94-	13	1.89022	5.09E-05	0.657695	0.028077	10.7342	0.018493	86.6713	neutral	1	0	-2.2343	14

5) <u>Consolidating time intervals and eliminating gaps in the transcript_data tables</u> –

In the 'transcript_data' tables, the 'start' and 'end' columns represent the time intervals for transcript scores.
To address time gaps, we calculate the mean between the previous 'end' and next 'start' values, replacing both 'end' and 'start' with this mean. We also round off the "end" column value of the last row in each table.
The tables are named 1_updated.csv, 2_updated.csv etc. for each of the candidates in the transcript_data folder.

6) <u>Overall Data table formed for each of the candidates</u> –

- The merged_emotion tables are merged with the transcript_data tables, to form overall_data_1.csv, overall_data_2.csv etc. for each of the candidates in the overall_data folder.
- We traverse through the transcript data table. For each row, the time interval is obtained from the "start" and "end" columns.
- Now, we traverse through the merged_emotion table and find all the rows for which "elapsed_time" column value falls within the time interval. Let the number of the rows be "n".
- The data in multiple rows is condensed to one row by taking mean of the numeric column values for these "n" rows in the merged_emotion table.
- This one row is now merged with the corresponding transcript_data table row.
- If no "elapsed_time" value is in the time interval of any row in transcript_data table, then null value is provided for the merged_emotion columns.
- Dominant and the next dominant emotion for each row is added in the overall_data table under "dominant_emotion" and "subsequent_emotion" columns.

Overall_data table

movie_id	start	end	angry	disgust	fear	happy	sad	surprise	neutral_emotion	gaze	blink	eye_offset	dominant_emo	subsequent_emo
93663f94-	0	5.56												
93663f94-	5.56	9.6	18.113	0.6831	18.02	2.8447	7.46	1.37268	51.50365	0.5	0	14.3835	neutral	angry
93663f94-	9.6	14.48	19.851	0.001	1.658	2.4541	8.01	2.29591	65.7335	0.5	0	15.15415	neutral	angry
93663f94-	14.48	18.48	16.46	0.0035	7.055	18.912	0.84	27.3113	29.41667667	0.33	0	18.65113	neutral	surprise

text	positive	negative	neutral_se	confident	hesitant	concise	enthusiast	speech_speed
Hello, I an	0.580265	0.152281	0.267454	0.846701	0.845698	0.635805	0.647783	2.517986
IIM Coiko	0.550327	0.189263	0.26041	0.679283	0.733701	0.544145	0.41739	3.217822
Technolog	0.63986	0.11115	0.24899	0.902729	0.83462	0.715861	0.700062	2.868852
of three y	0.441894	0.399186	0.158919	0.774308	0.813044	0.522462	0.279916	3.75

Text Analysis

(Codes for Text Analysis)

1) Cleaning of Transcript texts:
   The transcripts of the videos for all the candidates is provided. Tokenization, stop word removal, lowercasing and lemmatization performed on these text files and the cleaned text is stored in processed_text folder.

2) Relevant education, skills, experience and reasons to join derived from the cleaned texts.

3) Separation of "education", "skills", "experience" and "reasons to join" from the "text" column in overall_data_i.csv files:

   - Tokenization, stop word removal, lowercasing and lemmatization performed on the "text" column of the overall_data_i.csv files. The output is stored in a new column "cleaned_text".

Column names	Keywords
education	"postgraduate", "management", "iim", "degree", "bba", "mba", "varanasi university","engineering","graduate", "data science","python","data analysis","mass media","entrepreneurship","analytics","commerce"
skills	"attention","analyze", "detail", "consistency", "perseverance", "flexible attitude", "passionate", "passion","consulting","adaptive","inquisitive","learner","public relation","strategy","planning","writer","editor","analytical skill","finance","analytical","entrepreneurship","skill"
experience	"regulatory affair", "risk management", "three years", "investment", "framework venture", "startup business", "business model", "valuation", "finance model", "welfare society", "sale associate","advisor","administration","two year","consulting experience", "validation process software","insurance","underwriting","statutory audit","ca","audit","startup","leading project","business development","accounting associate","tax associate"
reasons_to_join	"looking challenging role", "skill practice", "believe", "fit", "grow", "acceleration","rewarding experience", "company flourish","aspiring result","share idea","program strategy", "spread awaremess", "positive attitude","affinity","strategizing","consultant","love learning","want join","newer experience","create society","child live","uniqueness","resonated","part","awakening","reasons to join": "edtech space","believe","fascinated","expand","successful model","aim","noble cause", "challenging","development","superior","best outcome"

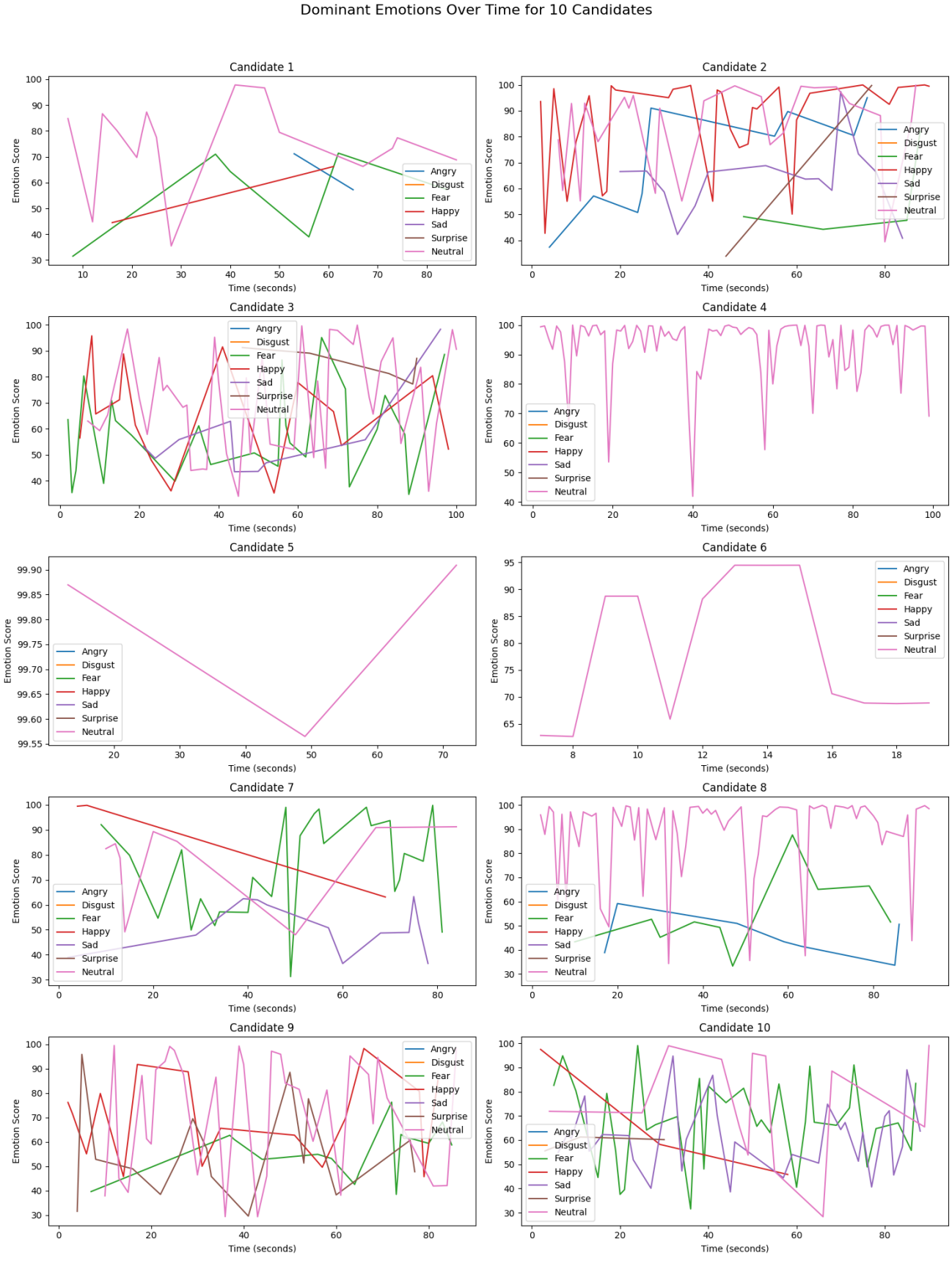
   - New columns "education", "skills", "experience" and "resaons_to_join" formed in overall_data_i.csv files
   - 1 or 0 is input in the new columns on the basis of the presence of these specific keywords with the corresponding column(education, skills, experience, reasons_to_join)

Data Visualizations and Insights

(Codes for Data Visualizations)

From merged_emotion_i.csv files

1. Dominant emotion of all the candidates over time



Insights obtained:

Candidate Number	Insights
1	Neutral and happy emotions are predominant with undertones of fear
2	Happy and neutral emotions are predominant with undertones of anger
3	Most fluctuating emotions. Neutral, happy, and fear emotions
4	Most consistent emotions. (neutral)
5	Insufficient emotion data
6	Insufficient emotion data. Neutral emotion shown
7	Fear emotion is predominant with happy emotion decreasing over time
8	Relatively consistent neutral emotions with slight undertones of fear
9	Neutral and happy emotions with undertones of surprise and fear
10	Fear and sad emotions fluctuating with happy emotion decreasing over time

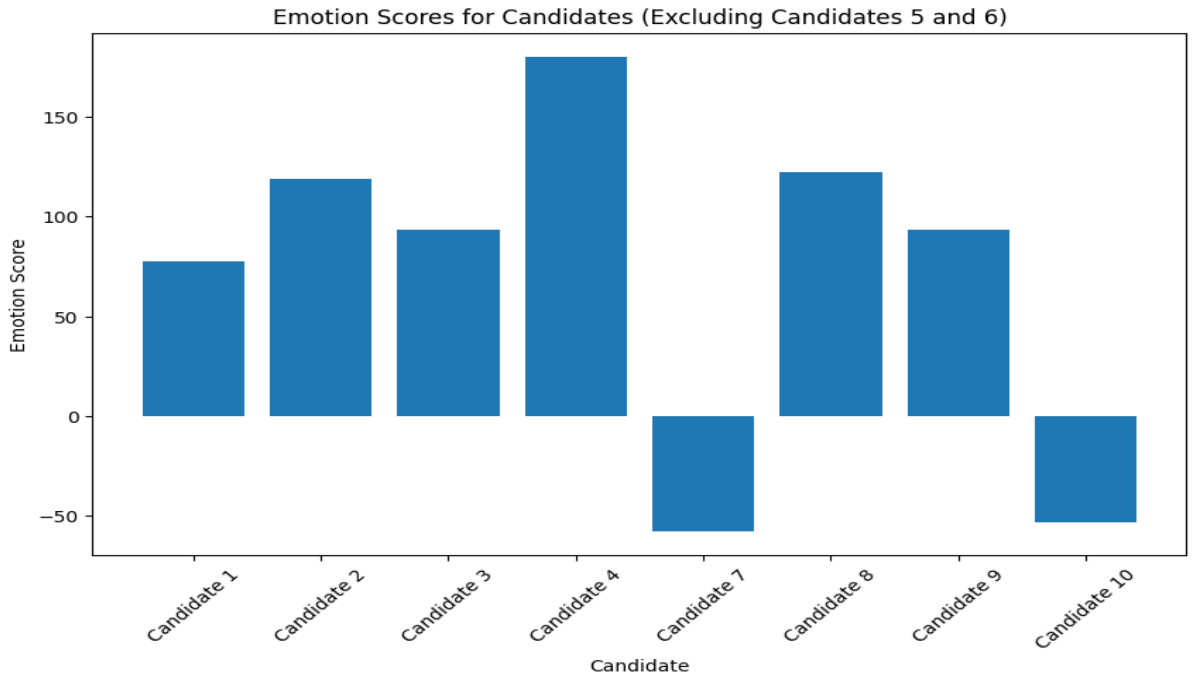
- Candidates 1, 2, 4, 8 and 9 show positive or neutral emotional response, while candidates 7 and 10 show negative emotional response over time.

2. Compare the emotion scores of all the candidates

Average of all the emotions of each of the candidates calculated.

Emotion scores: 'angry': -1, 'disgust': -1, 'fear': -2, 'happy': 3, 'sad': -1, 'surprise': 1, 'neutral': 2

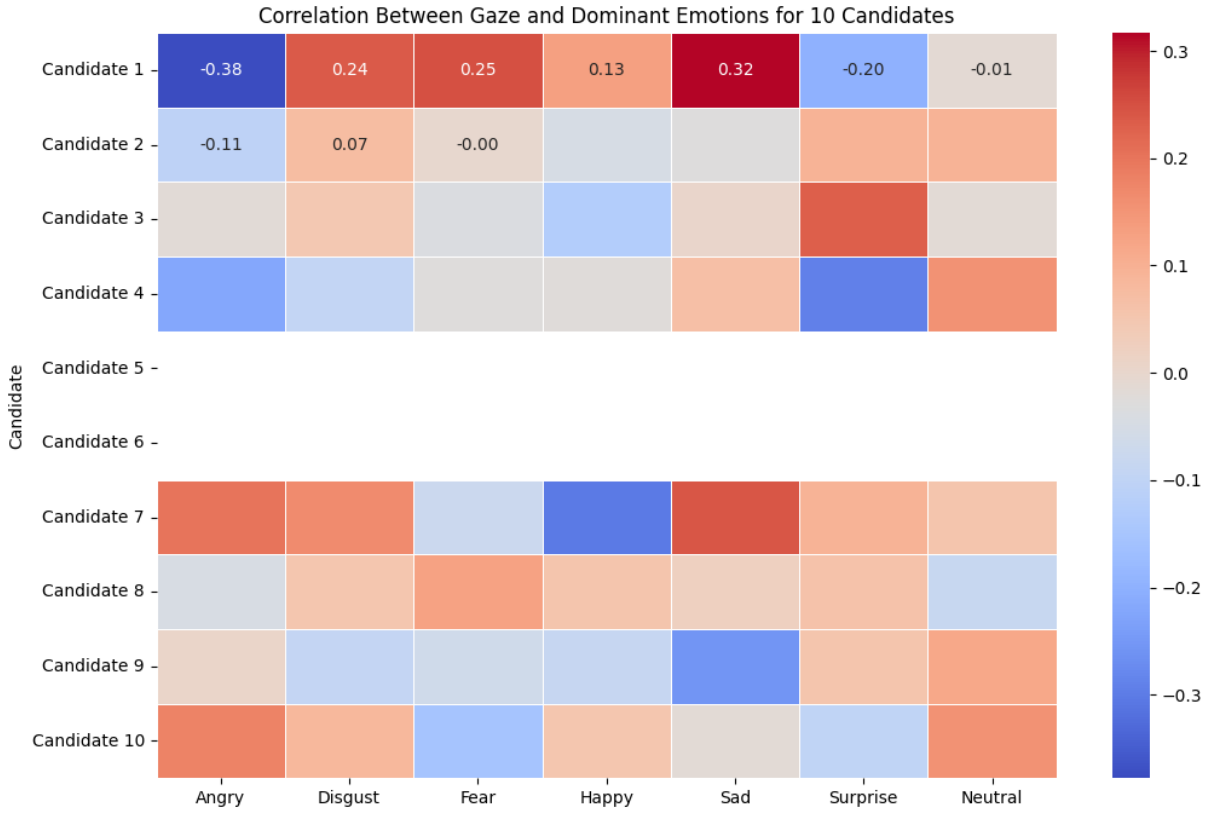
Total emotion score (for each candidate) = sum of (emotion multiplied by emotion score)



Insights obtained:

- Candidates 7 and 10 exhibit negative emotion scores
- Candidate 4 shows the highest emotion score
- Candidates 8, 2, 9 and 3 show good emotion scores

3. Correlation between gaze and dominant emotion

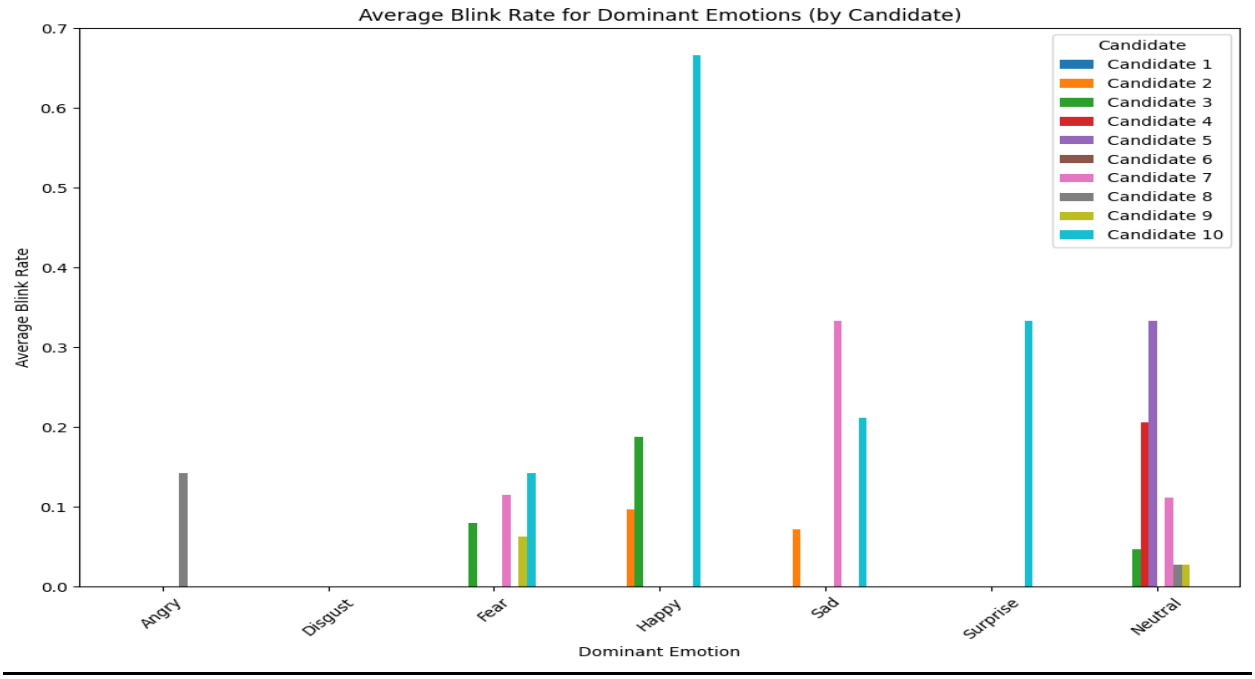


Insights obtained:

Candidate Number	Dominant emotions when looking at the camera		
	More	**Less**	**No Relation**
1	Sad, fear, disgust, happy	Angry, surprise	neutral
2	Surprise, neutral, disgust	angry	Happy, sad, fear
3	Surprise, disgust	happy	Neutral, angry, sad, fear
4	Neutral, sad	Surprise, angry	Happy, fear
5	Insufficient emotion data		
6	Insufficient emotion data		
7	Sad, angry, disgust, surprise, neutral	Happy, fear	
8	Fear, happy, disgust, surprise		Sad, neutral, angry
9	Neutral, surprise	sad	Fear, happy, disgust, angry
10	Neutral, angry, disgust, happy	fear	Sad, surprise

- For almost all the candidates, the dominant emotions (happy, neutral, fear) shown over time do not share much correlation with gaze.
- Candidates 2, 8, 9 and 10 do not show much correlation between gaze and dominant emotions
- Candidate 1 shows strong correlation between gaze and dominant emotions
- Candidates 3, 4 and 7 show moderate correlation between gaze and dominant emotions

4. Comparison of blink rate with dominant emotions for all the candidates, using bar charts

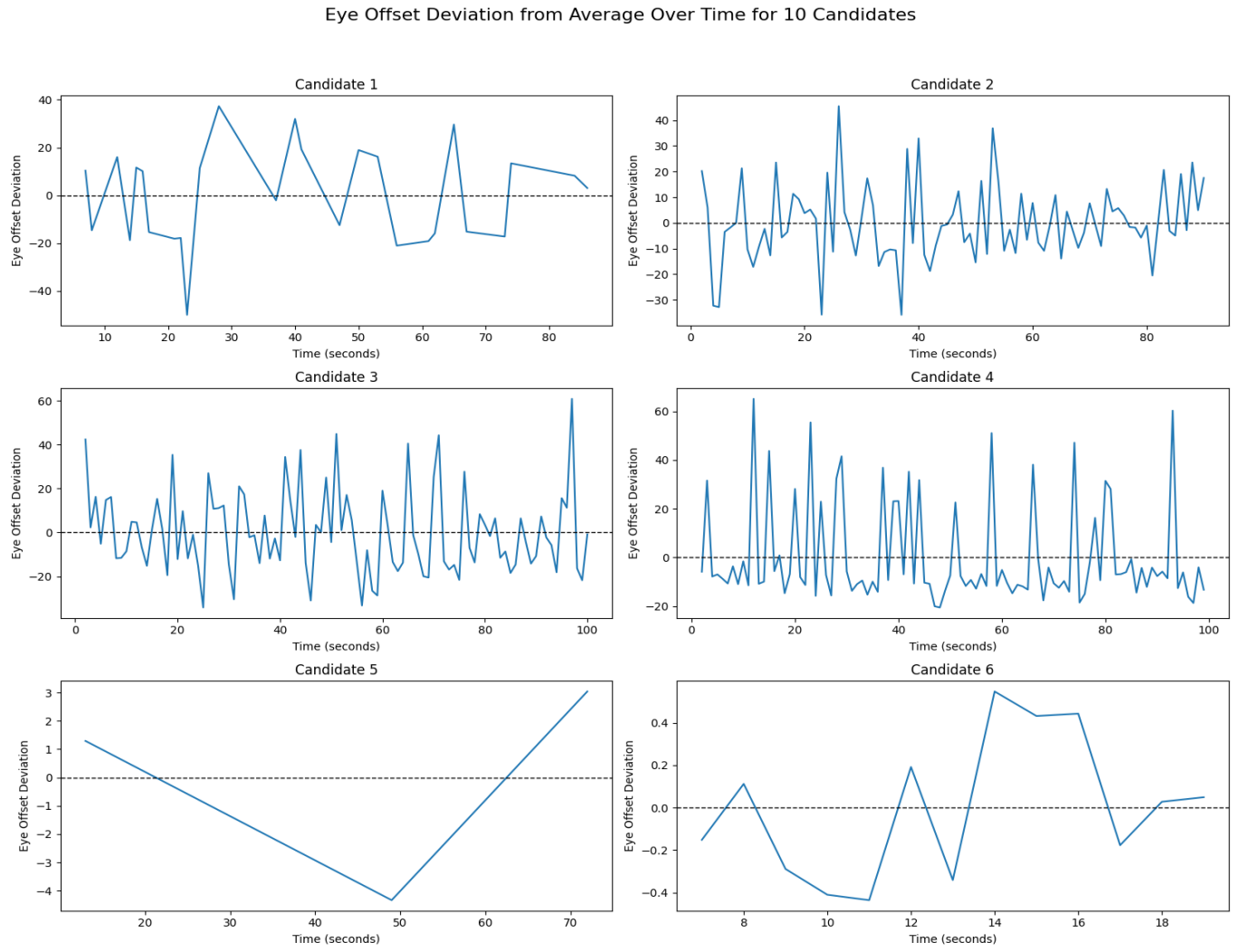


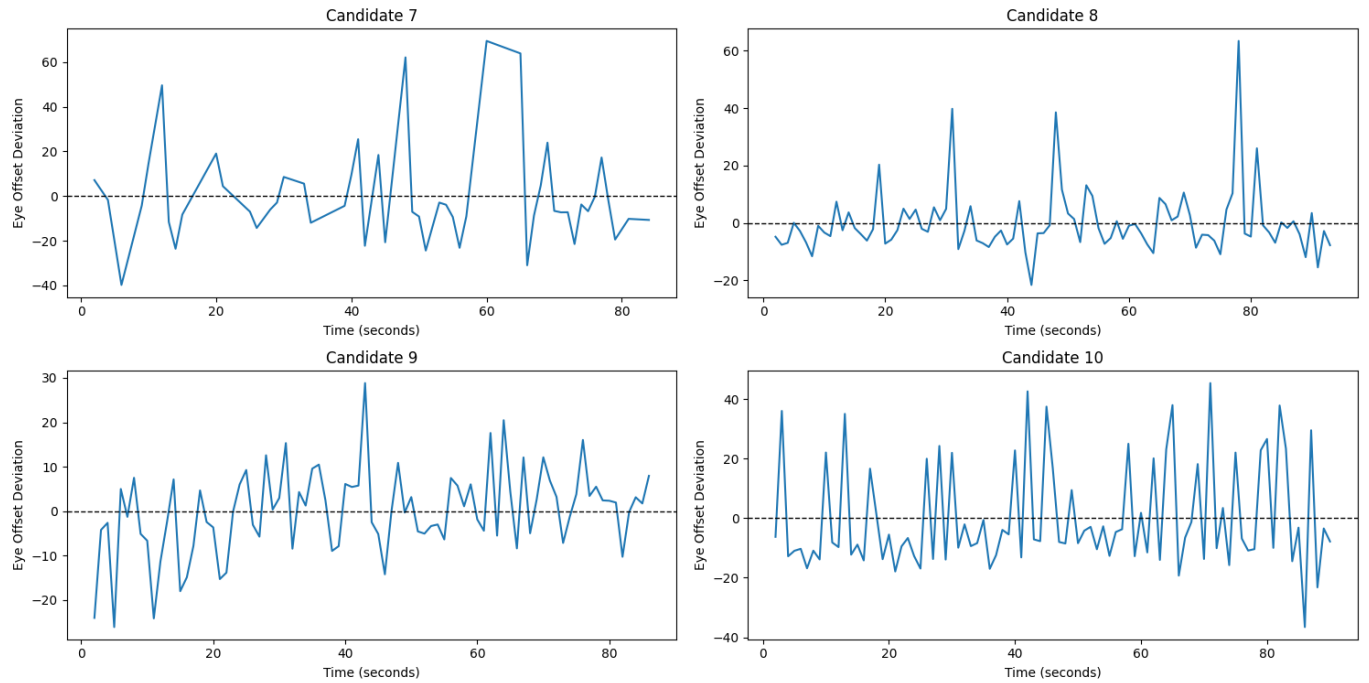
Insights obtained:

Candidate Number	Emotions when blink rate high
1	
2	Happy, sad
3	Happy, fear
4	Neutral
5	Neutral
6	
7	Sad, neutral, fear
8	Angry
9	fear
10	Happy, surprise, sad, fear

- Happy, neutral, sad and fear emotions are associated with higher blink rates.
- Candidates 7 and 10 show higher than average blink rate and also exhibit negative emotional response over time (insight as gained above)

5. Visualize the deviation of the eye offset over time with respect to the average eye offset for all the candidates

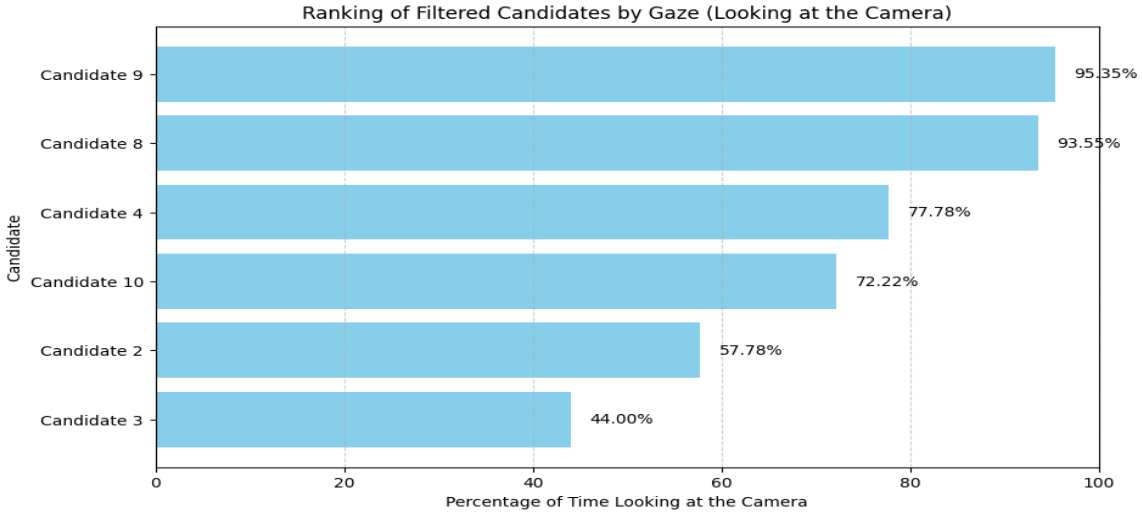




Insights obtained:

- Candidates with Highest to Lowest mean Eye Offset Deviation from average eye offset (of each candidate) over time:
  - Candidate 1: 17.60
  - Candidate 7: 15.82
  - Candidate 4: 15.64
  - Candidate 3: 14.90
  - Candidate 10: 14.57
  - Candidate 2: 11.58
  - Candidate 9: 7.22
  - Candidate 8: 7.09
  - Candidate 5: 2.89
  - Candidate 6: 0.28

6. **Bar chart to compare the percentage of time spent by the candidates looking at the camera. (candidates with total number of rows greater than the median number of rows of all candidates)**

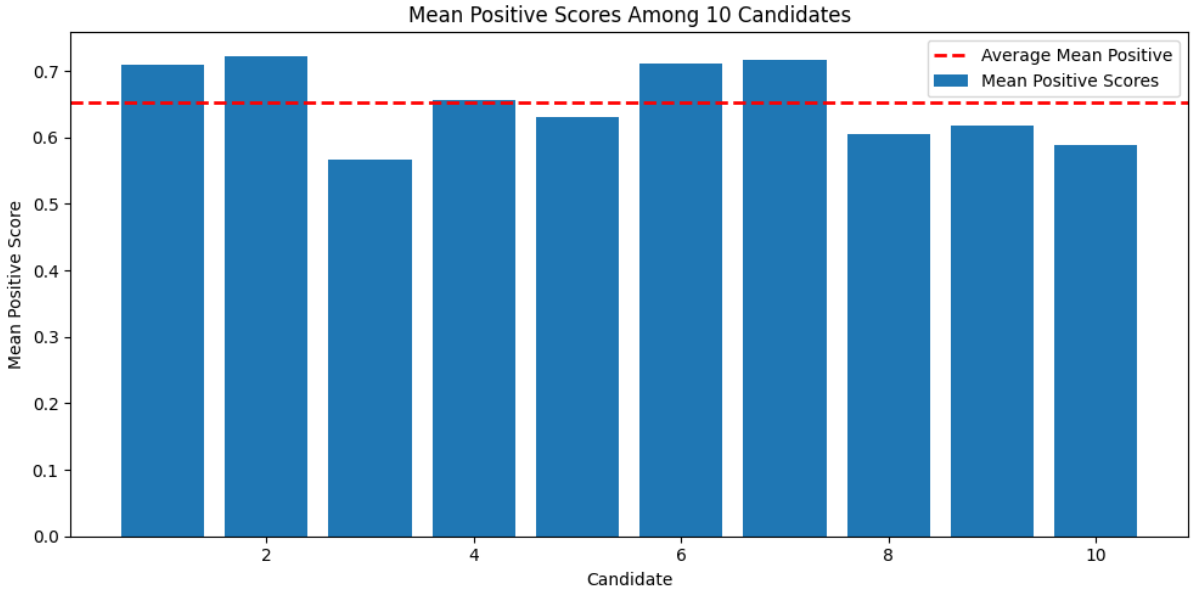


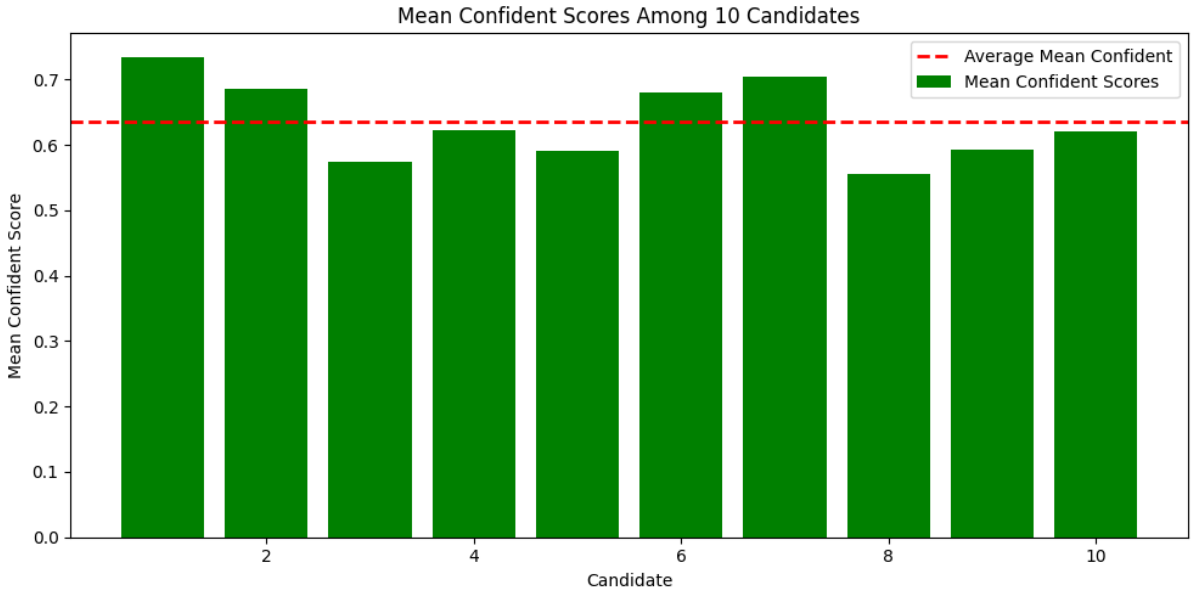
Insights obtained:

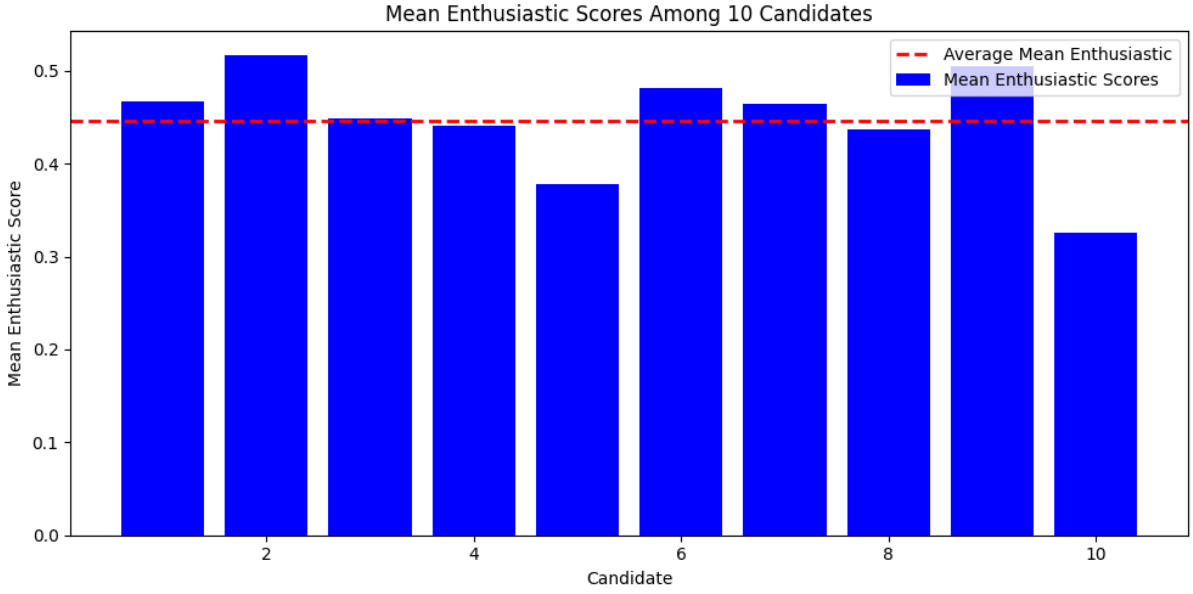
- Candidates 9 and 8 show the most engagement with the camera.
- Candidates 2 and 3 show the least engagement to the camera.
- Candidates 4 and 10 show above average engagement with the camera.

From Transcript_data tables

1. **Comparison of mean positive, confident and enthusiastic scores among all the candidates**







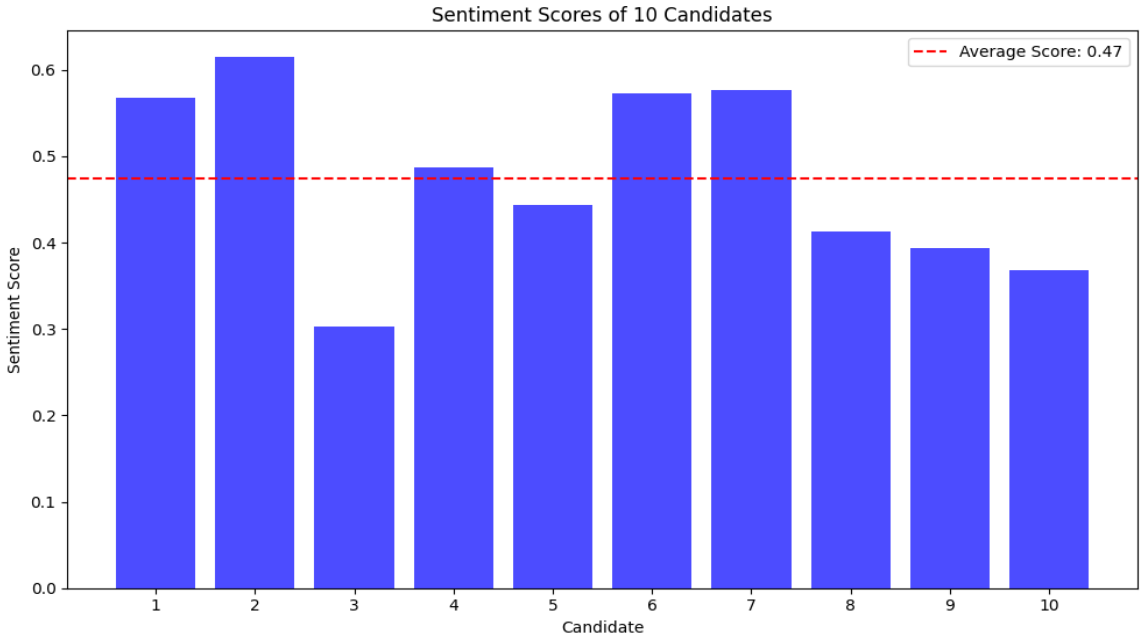
Insights obtained:

- Candidates 1, 2, 6 and 7 show higher than average positive, enthusiastic and confident scores.
- Candidates 3, 5, 8 and 10 show lower than or equal to average positive, enthusiastic and confident scores.
- Candidate 9 shows high enthusiastic score, while candidates 10 and 5 show lack of enthusiasm as compared to other candidates.

Candidate Number	Comparison with average		
	Positive	Confident	Enthusiastic
1	Above	Above	Above
2	Above	Above	Above
3	Below	Below	Average
4	Average	Average	Average
5	Average	Below	Below
6	Above	Above	Above
7	Above	Above	Above
8	Below	Below	Average
9	Below	Below	Above
10	Below	Average	Below

2. Comparing the overall sentiment scores of all the candidates

Average of positive, negative and neutral of each of the candidates calculated.
Scores: 'negative': -1, 'positive': 1, 'neutral': 0
Overall sentiment score (for each candidate) = sum of (score*average value)

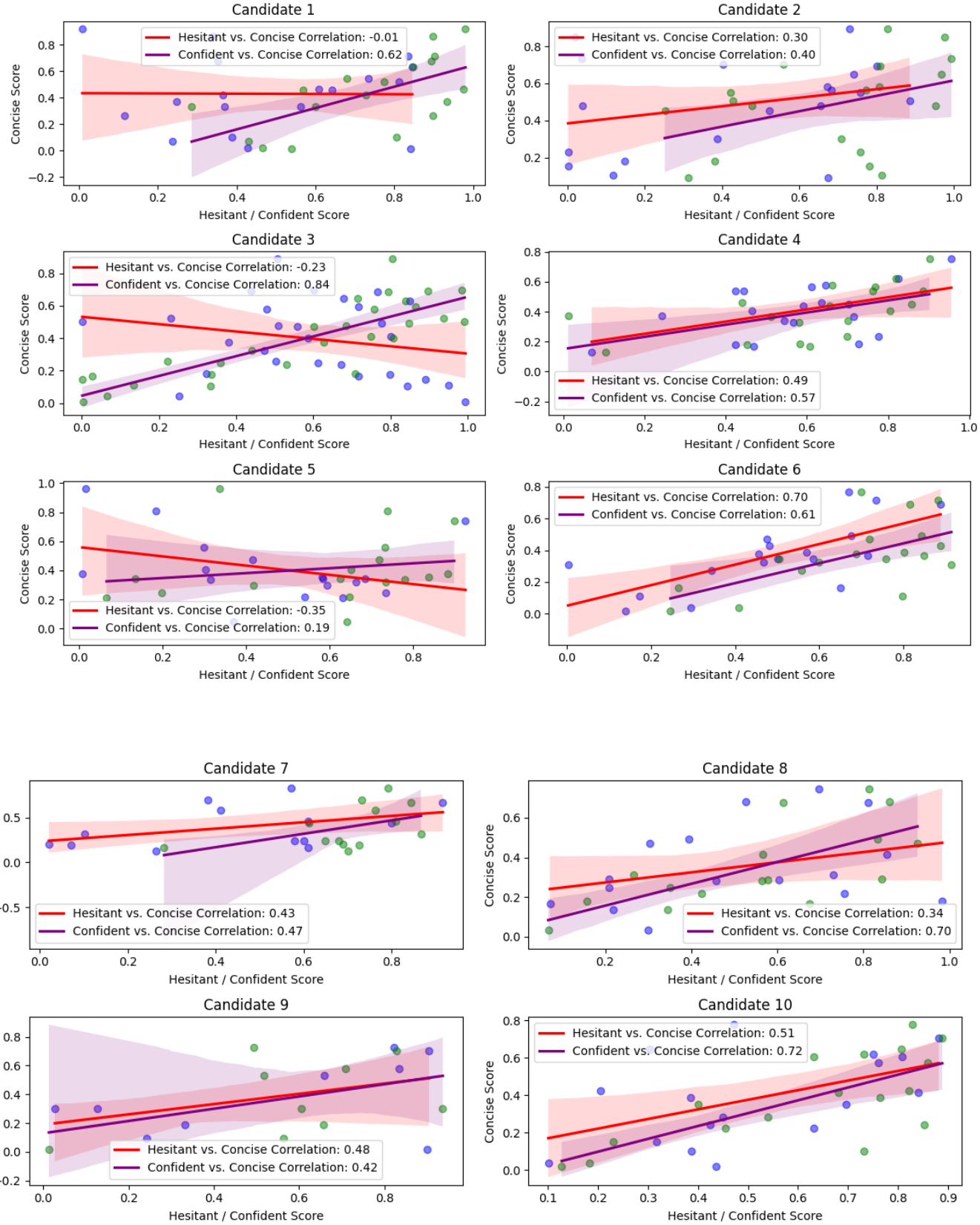


Insights obtained:

- Candidates 2, 7, 6 and 1 show high sentiment scores
- Candidates 4 and 5 show near average sentiment scores
- Candidates 8, 9, 10 and 3 show low sentiment scores

3. Hesitant vs Concise and Confident vs Concise correlations for all the candidates

Correlation Between Hesitant and Concise, and Confident and Concise Scores for 10 Candidates

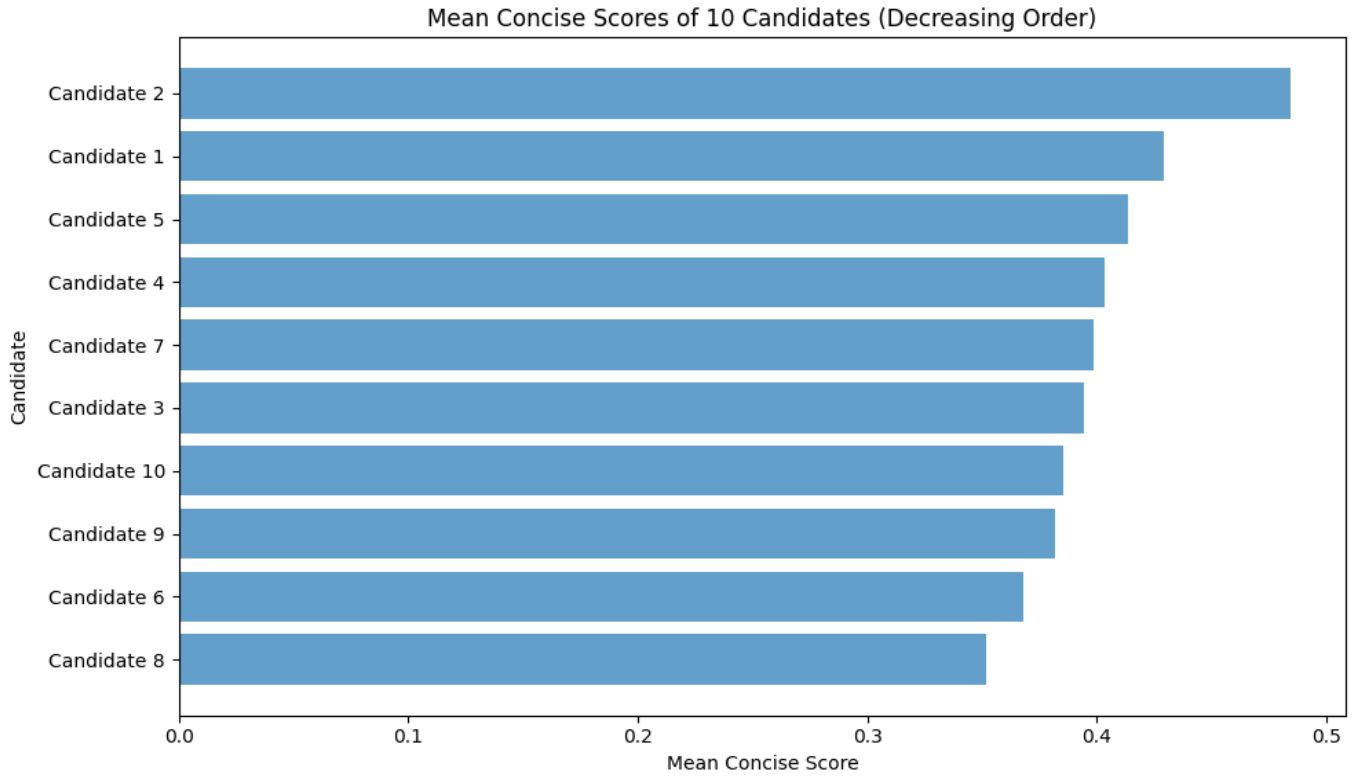


Insights obtained:

Candidate Number	More hesitant, less concise	More confident, more concise
1	Yes	Yes
2	No	Yes
3	Yes	Yes
4	No	Yes
5	Yes	Yes
6	No	Yes
7	No	Yes
8	No	Yes
9	No	Yes
10	No	Yes

- For candidates 1, 3 and 5, their concise scores are dependent on the level of confidence.
- For all other candidates, the concise score is independent of their level of confidence.

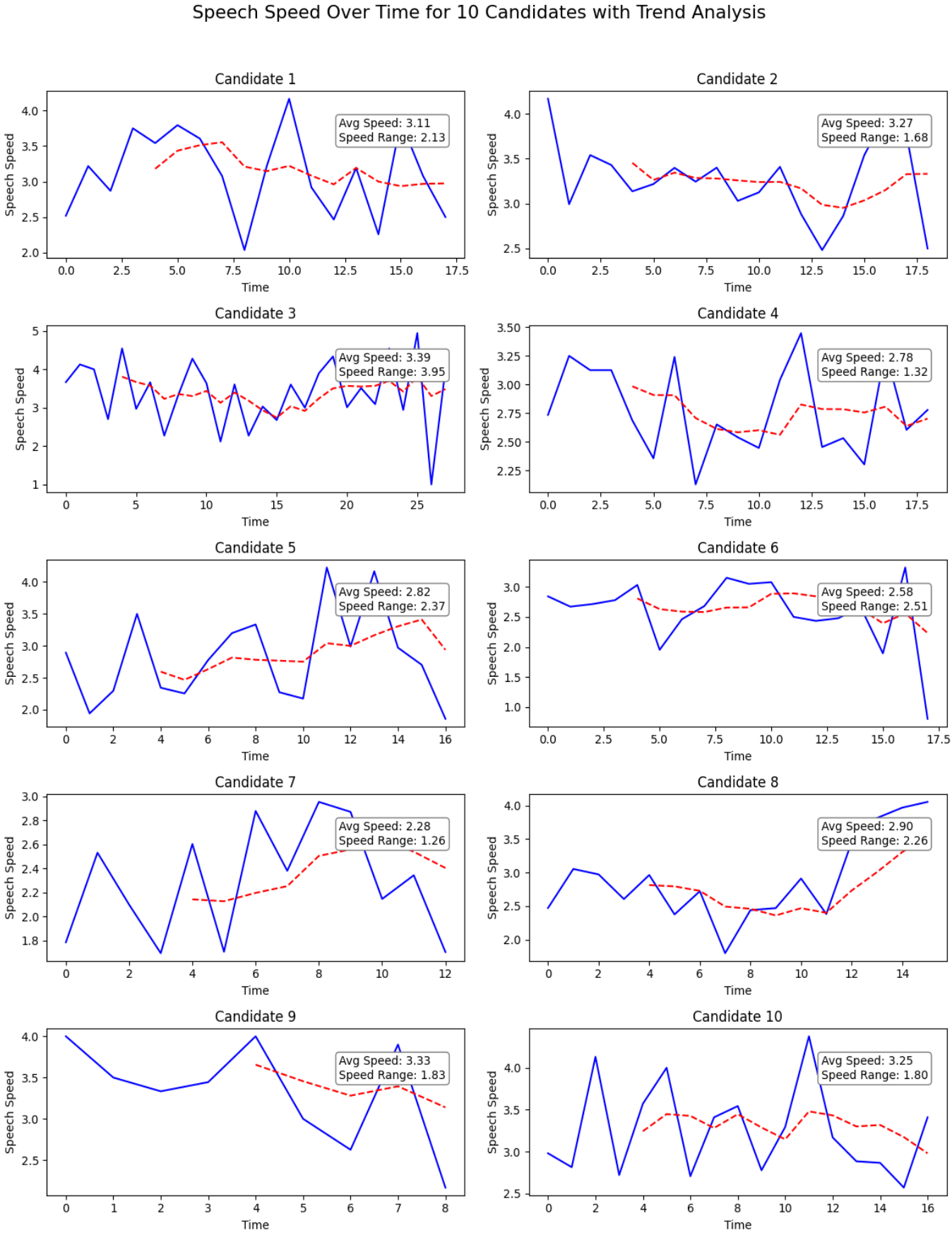
4. Comparison of concise score of all the candidates



Insights obtained:

- Candidates 2, 1, and 5 have high concise scores
- Candidates 8, 6, 9 and 10 have low concise scores

5. Speech speed of candidates over time



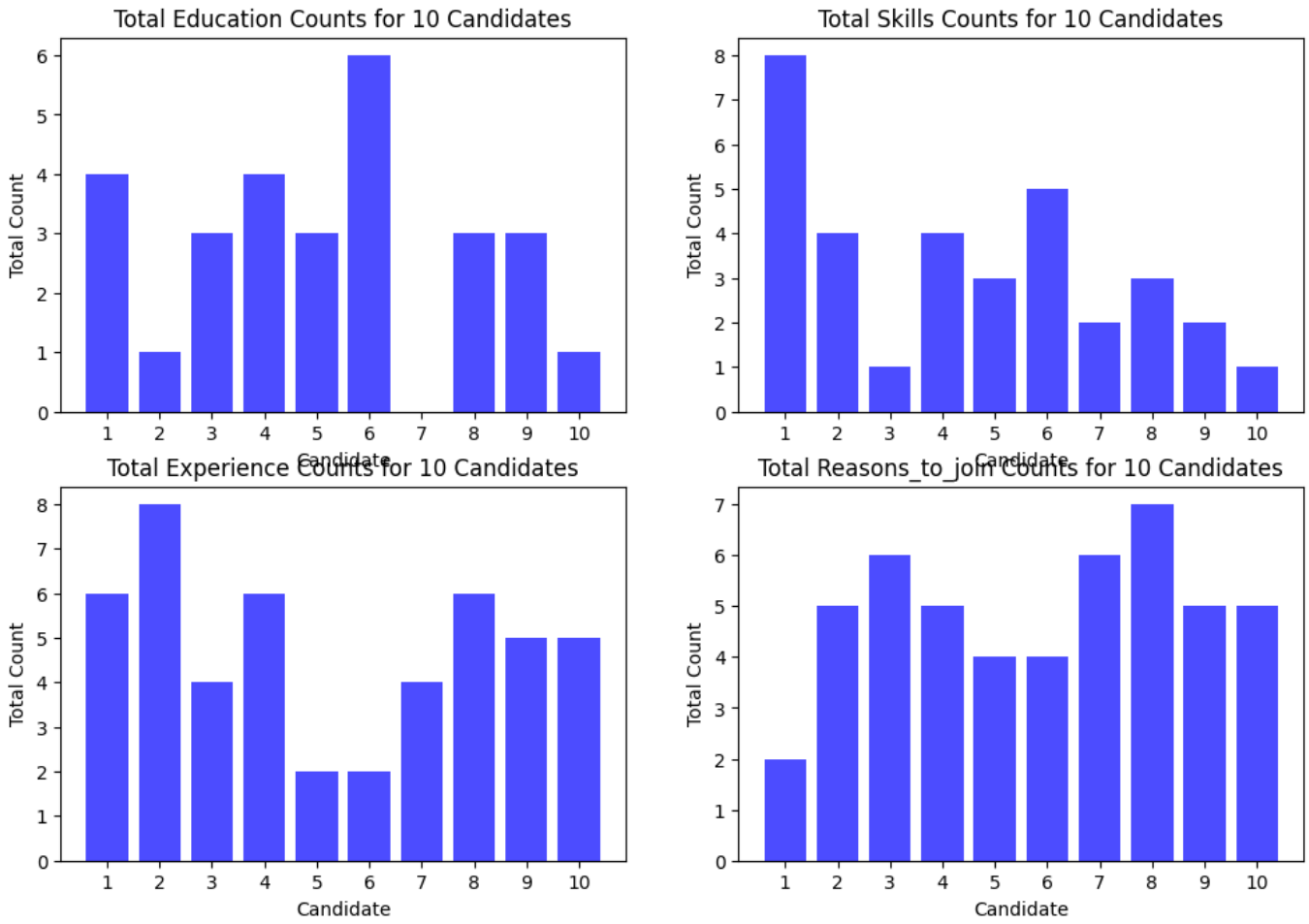
Insights obtained:

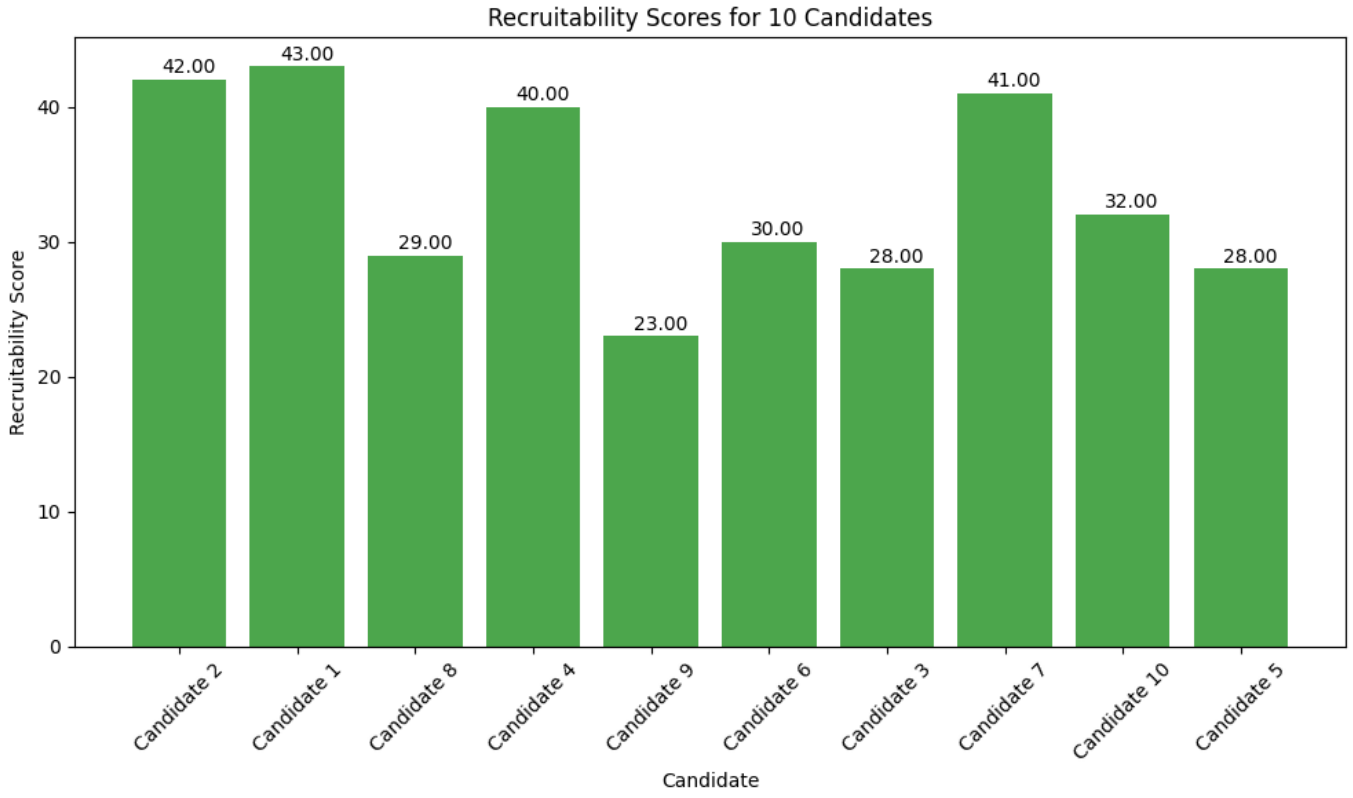
• Candidates 3 shows high speech speed fluctuations in smaller intervals of time. • Candidate 9, 7 and 4 have consistent pacing in their speeches based on speech speed range
• Candidates 1 and 9 show decrease in speech speed over time. • Candidates 5, 7 and 8 show increase in speech speed over time.
• Candidates 3, 9, 2 and 10 have a high average speech speed • Candidates 7, 6 and 4 have lower speech speeds.

From Overall Data tables:

1. **Plotting charts on the basis of the relevancy of education, skills, experience and reasons to join, and finding the recruit_ability scores for each of the candidates**

   Calculating recruit_ability score:
   Scores: education – 1, skills – 2, experience – 3, reasons to join – 2
   Recruit_ability score = sum of (scores*number of relevant education, skills etc.)



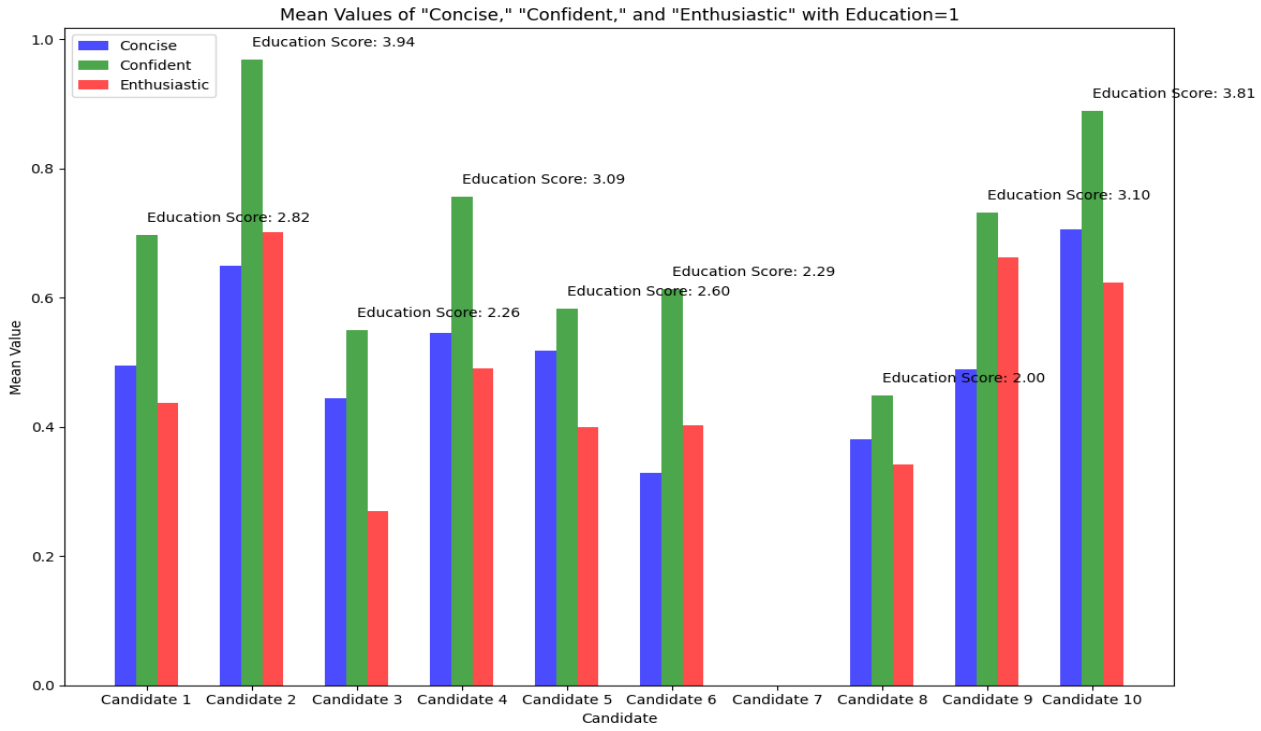


Insights Obtained:

- Candidate 6 with most education count
- Candidate 1 with most skills count
- Candidate 2 with most experience count
- Candidate 8 with most reasons to join count
- Ability to be recruited in decreasing order of recruit_ability score – 1, 2, 7, 4, 6, 8, 3, 5, 9

2. **Finding "education_sentiment" when all the candidates mention their education in their videos**

Education_sentiment score distribution – concise - 2, confident - 2, enthusiastic - 1



Insights obtained:

Candidates in the decreasing order of emotion_sentiment scores – 2, 10, 9, 4, 1, 5, 6, 3

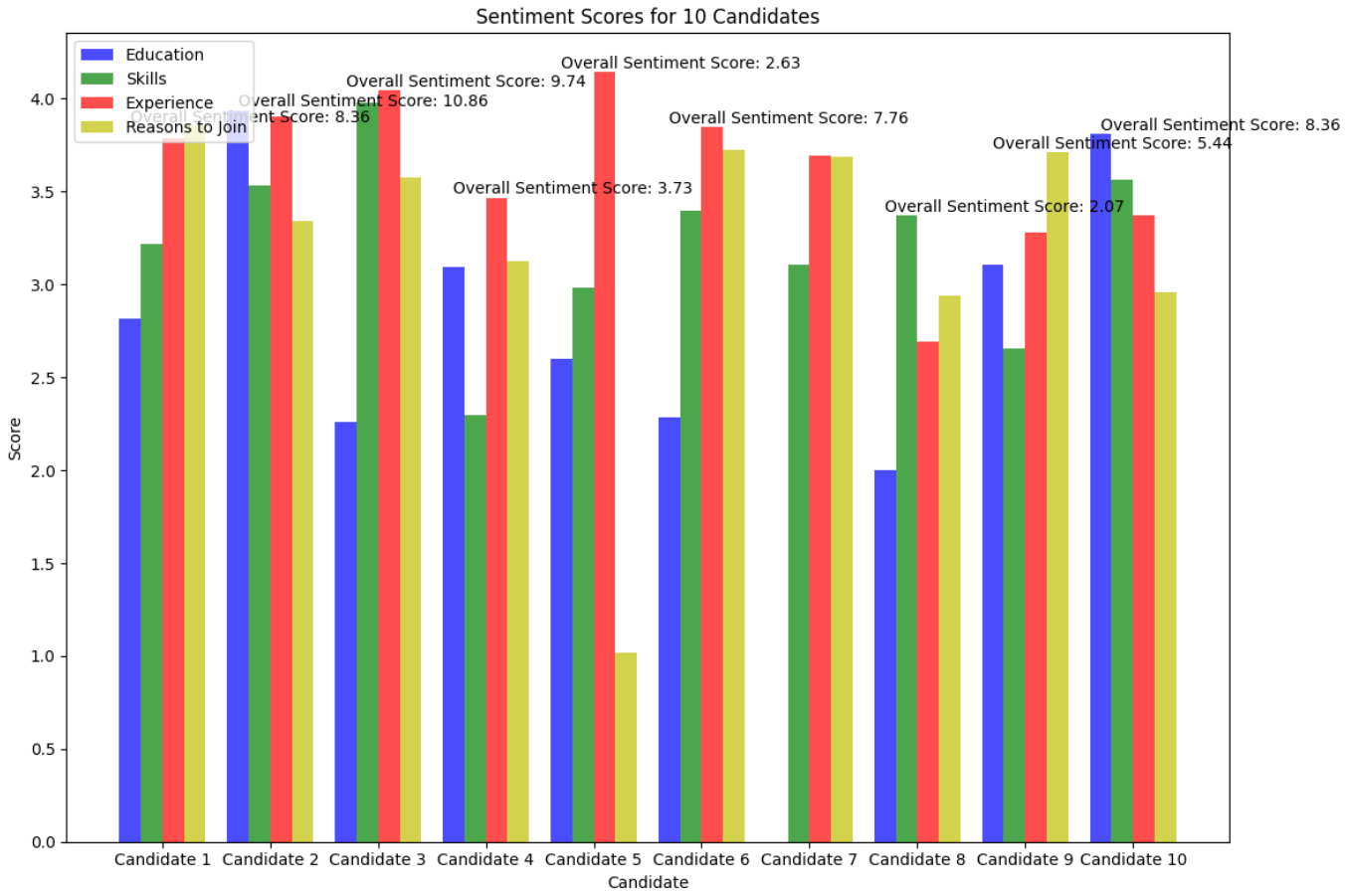
3. **Similarly, "skills_sentiment" calculated for all the candidates:**
skills_sentiment score distribution – concise - 2, confident - 3, enthusiastic – 1

Also, experience_sentiment calculated:
(concise - 2, confident - 3, enthusiastic - 2)

And, reasons_to_join_sentiment calculated:
(concise - 1, confident - 2, enthusiastic - 3) for each of the candidates.

4. **Finding "overall_sentiment_score" for each of the candidates and plotting the four types of sentiment scores**

overall_sentiment_score calculation:
(education_score - 1, skills_sentiment - 2, experience_sentiment - 4, reasons_to_join_sentiment - 3)

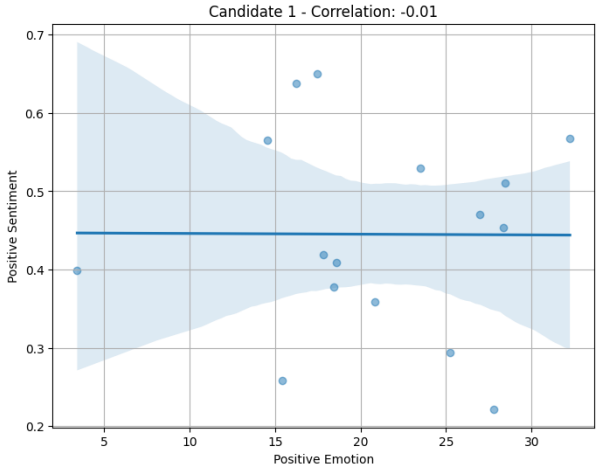


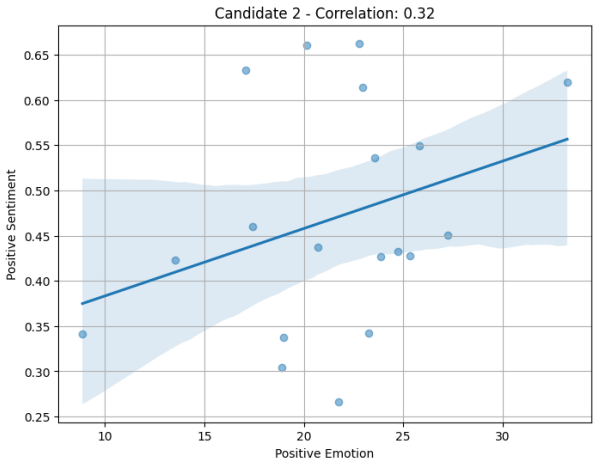
Insights obtained:

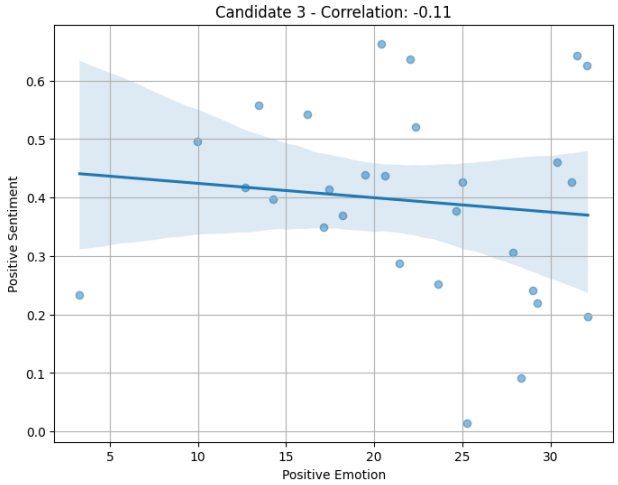
Candidates in the <u>decreasing order</u> of overall_sentiment_scores – 2, 3, 1=10, 6, 9, 4, 5, 8

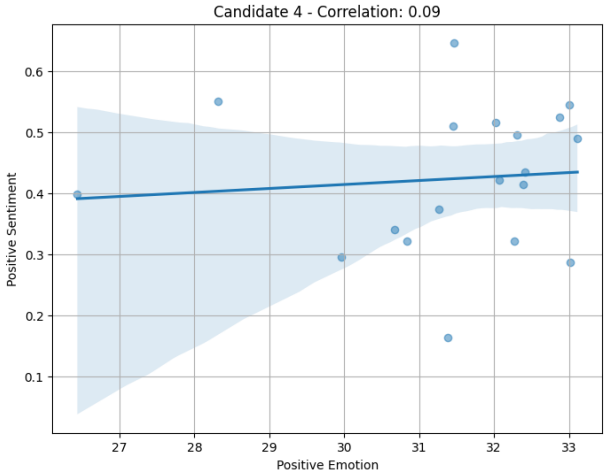
5. **Finding the correlation between "positive emotion" and "positive sentiment" for all the candidates using pearson correlation coefficient**

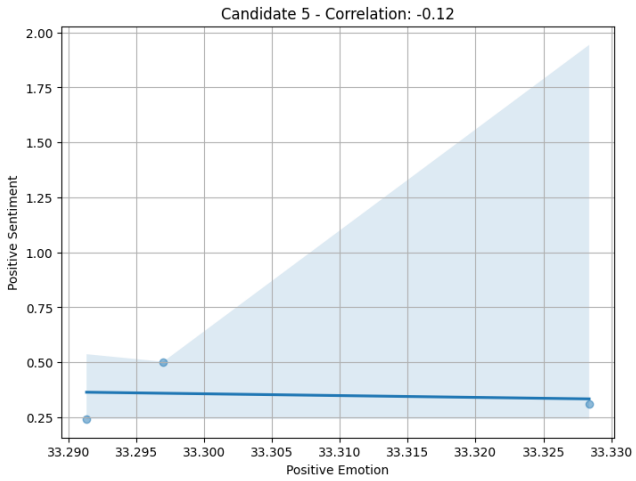
Positive emotion: mean of "happy", "surprise" and "neutral_emotion" value for each row
Positive sentiment: mean of "postive", "neutral_sentiment" and "enthusiastic" value for each row

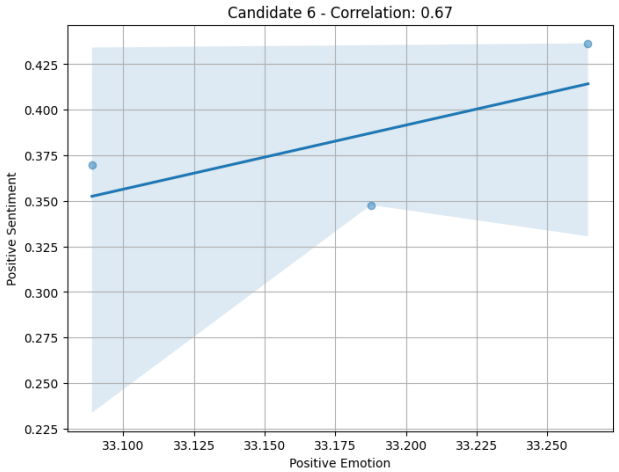


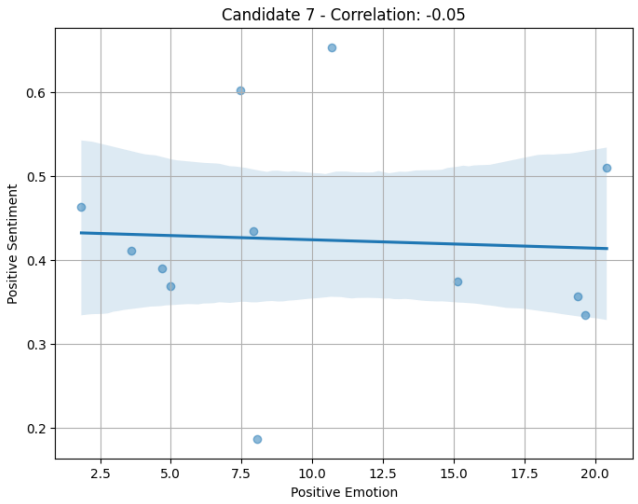


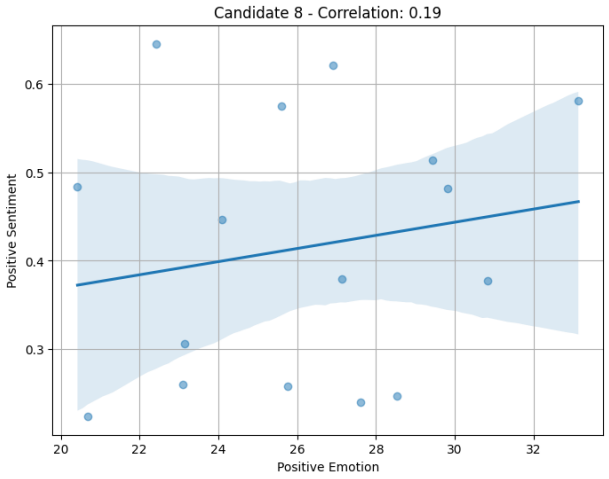


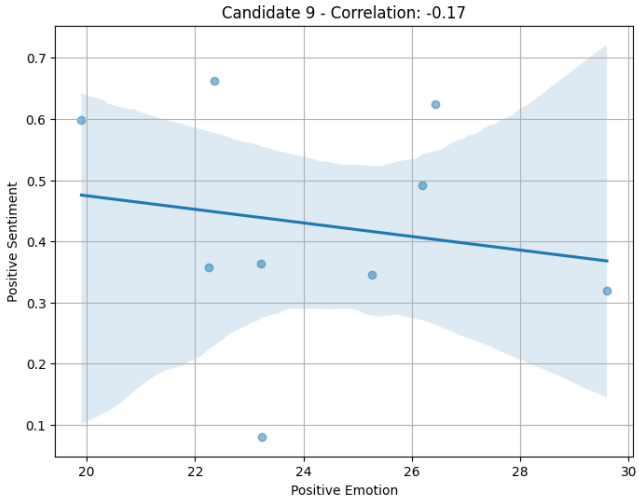


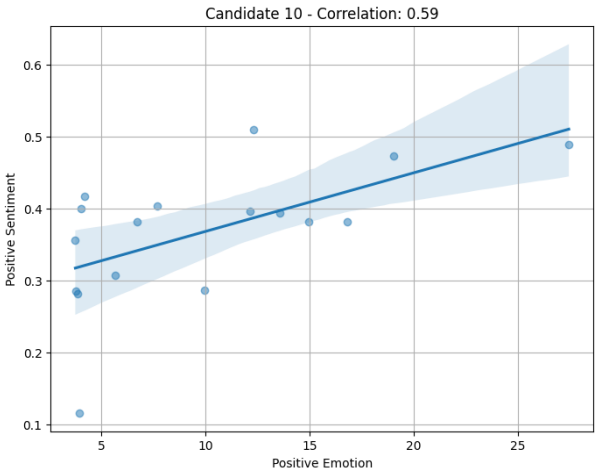












Insights Obtained:

```
Candidate 1 - Correlation: -0.01
Candidate 2 - Correlation: 0.32
Candidate 3 - Correlation: -0.11
Candidate 4 - Correlation: 0.09
Candidate 5 - Correlation: -0.12
Candidate 6 - Correlation: 0.67
Candidate 7 - Correlation: -0.05
Candidate 8 - Correlation: 0.19
Candidate 9 - Correlation: -0.17
Candidate 10 - Correlation: 0.59
```

Candidates 6 and 10 show high positive relation between “positive emotion” and “positive sentiment”

Candidates 8 and 2 show moderate positive relation between “positive emotion” and “positive sentiment”

Final Insights

Based on these insights, here's an analysis of the recruitment potential for each candidate:

Candidate 1:

- Shows positive or neutral emotional response.
- Strong correlation between gaze and dominant emotions.
- Higher than average positive, enthusiastic, and confident scores.
- Most skills count.
- High recruit_ability score.
- High overall sentiment score.

Recruitment Potential: Strong candidate for recruitment.

Candidate 2:

- Shows positive or neutral emotional response.
- Does not show much correlation between gaze and dominant emotions.
- High sentiment scores.
- Most experience counts.
- High recruit_ability score.
- High overall sentiment score.

Recruitment Potential: Strong candidate for recruitment.

Candidate 3:

- Good emotion scores.
- Moderate correlation between gaze and dominant emotions.
- Low sentiment score.
- High recruit_ability score.
- High overall sentiment score.

Recruitment Potential: Moderate candidate for recruitment.

Candidate 4:

- Shows positive or neutral emotional response.
- Above average engagement with the camera.
- High emotion score.
- High recruit_ability score.
- Lower overall sentiment score.

Recruitment Potential: Strong candidate for recruitment.

Candidate 5:

- Lower than average positive, enthusiastic, and confident scores.
- Near average sentiment scores.
- Low concise score.

Recruitment Potential: Weaker candidate, but not necessarily excluded.

Candidate 6:

- High positive, enthusiastic, and confident scores.
- Most education counts
- Low concise score.

Recruitment Potential: Reasonable candidate, but concise communication could be improved.

Candidate 7:

- Negative emotional response.
- Negative emotion scores.
- High blink rate.
- High sentiment scores.
- High recruit_ability score.

Recruitment Potential: Despite the negative emotional response, other factors suggest potential recruitment.

Candidate 8:

- Good emotion scores.
- Above average engagement with the camera.
- Low sentiment scores.
- Low concise score.

Recruitment Potential: Weaker candidate, but not necessarily excluded.

Candidate 9:

- Positive or neutral emotional response.
- Good emotion scores.
- High enthusiastic score.
- Low sentiment scores.
- High recruit_ability score.

Recruitment Potential: Reasonable candidate, with potential for improvement in sentiment.

Candidate 10:

- Negative emotional response.
- Negative emotion scores.
- High blink rate.

Recruitment Potential: Weaker candidate, negative emotions and blink rate may be concerns.