Data Science
Assignment Report

Sidharth Bajaj

Problem Statement

The assignment involves analyzing a dataset of 10 candidates, which includes emotion scores, transcript scores, and transcripts from their introduction videos. The main objectives are:

1. Determine candidate recruitability with supporting reasons from the data.
2. Analyze communication skills and identify areas of expertise based on the data.
3. Generate additional insights to aid in decision-making about the candidates.

The task requires preprocessing data, creating effective prompts, and performing Exploratory Data Analysis (EDA) to extract meaningful and actionable information.

Methodology

The analysis was conducted using Python, leveraging various libraries such as pandas, numpy, matplotlib, seaborn, and plotly. The approach can be broken down into several key steps:

1. Data Loading and Preprocessing
2. Primary Analysis
3. Secondary Analysis
4. Skillset and Other Data Analysis
5. Multiple Approaches for Job Suggestions

1. Data Loading and Preprocessing

- Emotion data and gaze data were loaded from CSV files for each candidate.
- Transcript data was also loaded and preprocessed.
- The data was organized into separate DataFrames for each candidate and type of data (emotion, gaze, transcript).

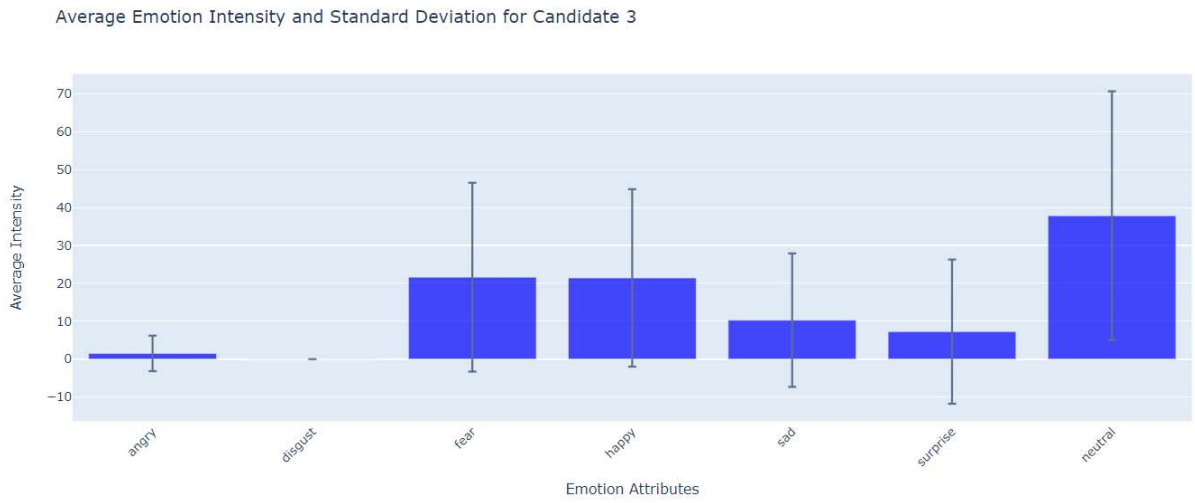
2. Primary Analysis

Emotion Data Analysis

- Calculated average emotion intensities for each candidate.
- Computed standard deviations to assess emotional variability.
- Visualized emotion data using bar plots with error bars to represent variability.

Threshold Count

Looking at the data, taking Average for further analysis was looking a valid approach but at the same time to measure the consistency of the candidates throughout the video, I had an idea which included setting a threshold to every emotion which would be same for the overall data and maintaining a count which would increase every time a candidate's emotion increased the threshold of that particular emotion.
For that time I dropped that idea as I didn't had a definite method to incorporate the same in the final score.

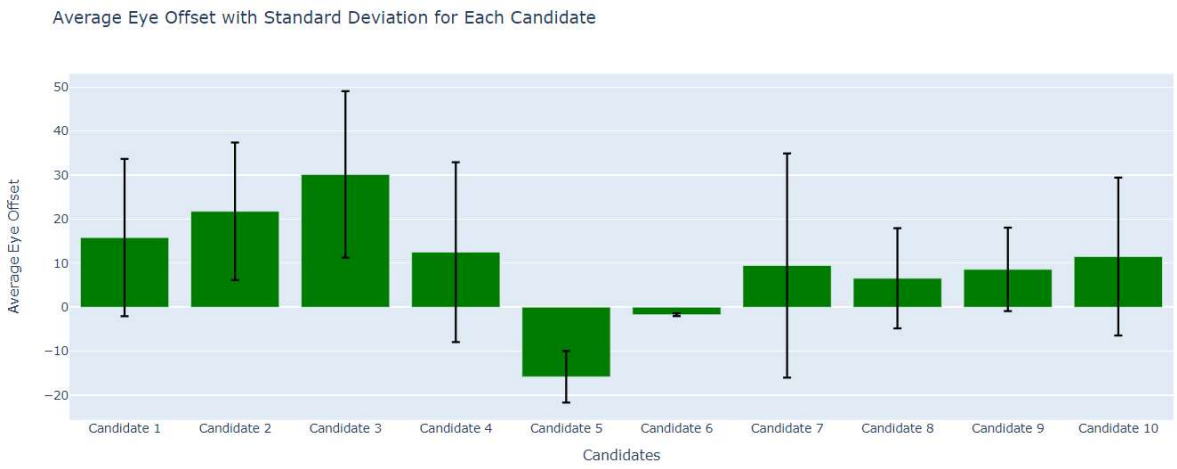


Constructed above plots for every candidate for data visualization.

Gaze Data Analysis

- Analyzed gaze patterns, eye offset, and blink data.
- Created visualizations to compare gaze-related metrics across candidates.

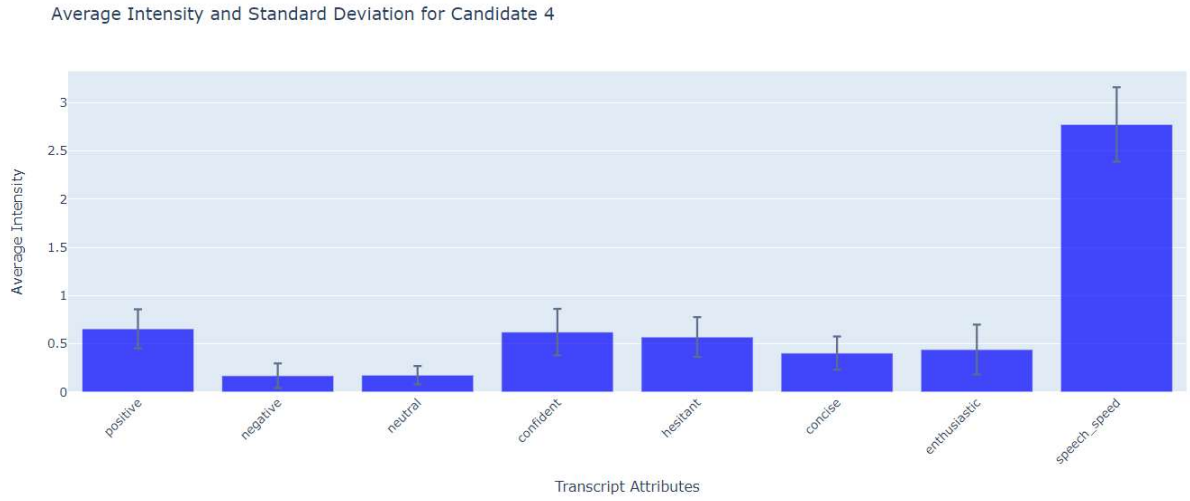
I



Constructed similar plots for gaze and blink data.

Transcript Data Analysis

- Examined transcript attributes such as positive/negative sentiment, confidence, hesitation, conciseness, enthusiasm, and speech speed.
- Visualized transcript data to identify patterns in communication style.



Final Scoring

Developed a weighted scoring system combining emotion, gaze, and transcript data.
Created a final score for each candidate to aid in ranking.
Visualized the final scores using a bar plot.

The weights were extracted by using Artificial Intelligence and Prompt Engineering. More on them is mentioned inside the Prompt Engineering Report.

```
weights = {
    'angry': -0.2,
    'disgust': -0.1,
    'fear': -0.1,
    'happy': 0.4,
    'sad': -0.2,
    'surprise': 0.1,
    'neutral_x': 0.1,
    'gaze': 0.05,
    'eye_offset': 0.05,
    'positive': 0.2,
    'negative': -0.1,
    'neutral_y': 0.1,
    'confident': 0.3,
    'hesitant': -0.1,
    'concise': 0.1,
    'enthusiastic': 0.3,
    'speech_speed': 0.1
}
```

The above are the weights assigned to every data we have and on the basis of these weights I calculated a function. I used the data extracted from primary analysis on this function and calculated a final overall score for each candidate.

The plot representing the final score:

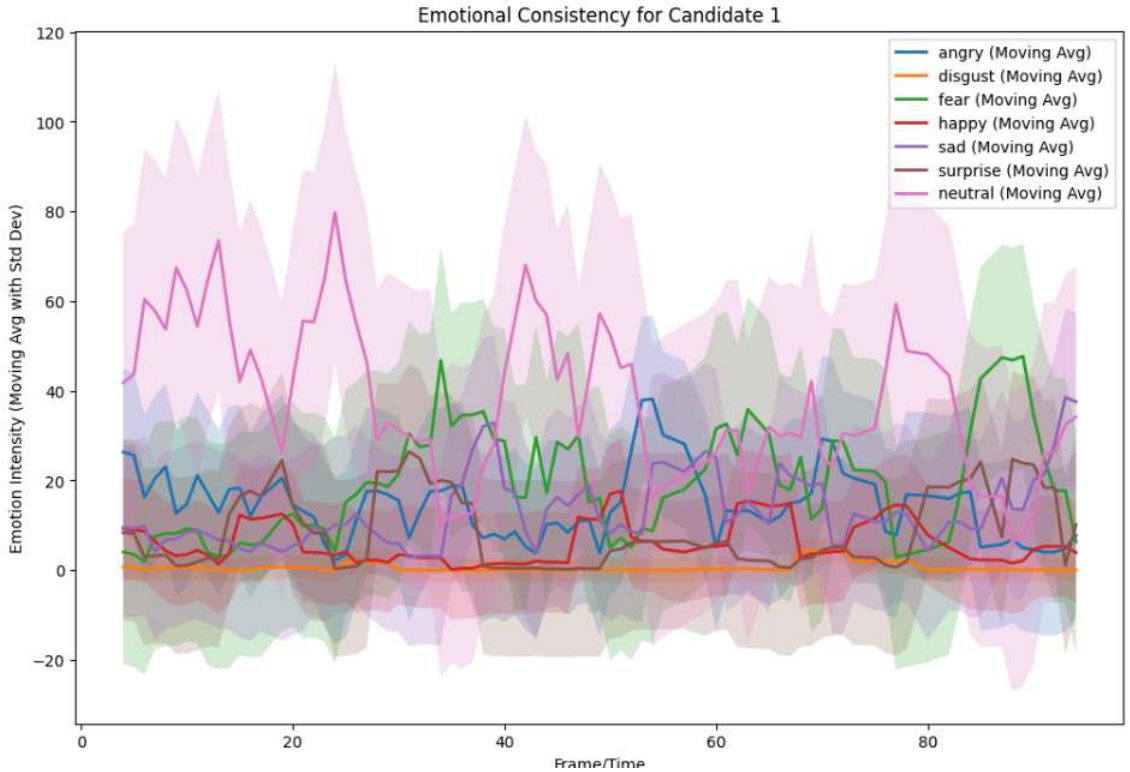
3. Secondary Analysis

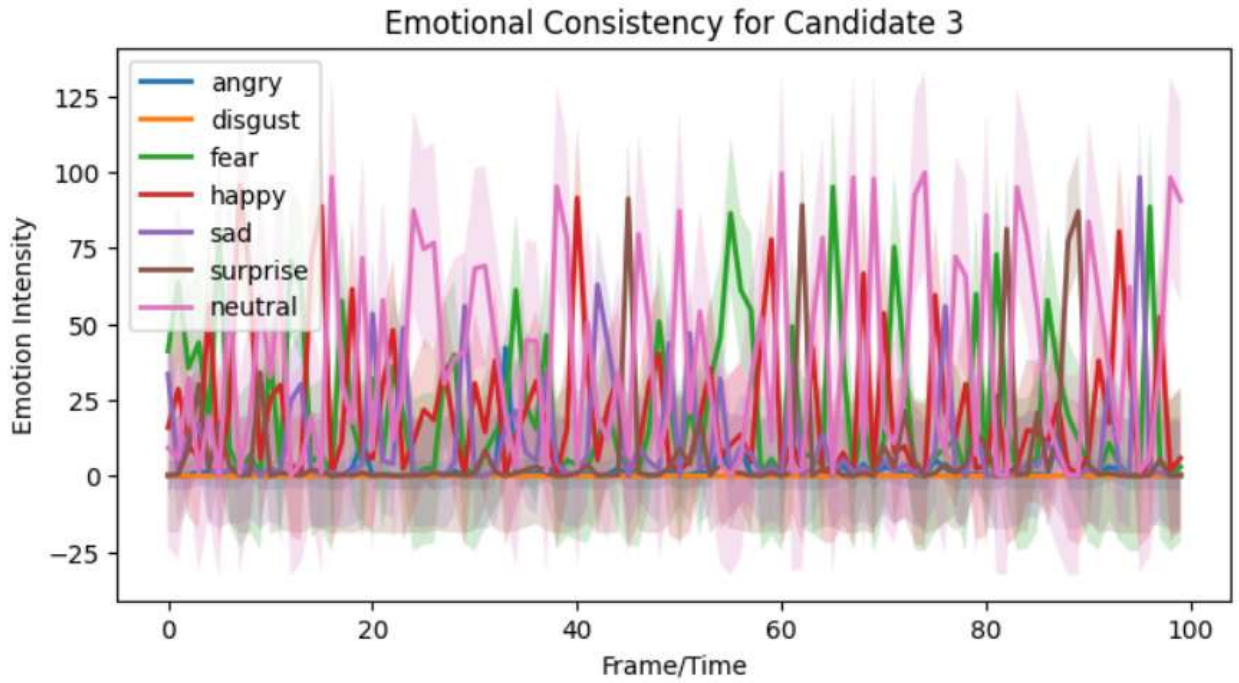
There was a significant problem with the primary analysis. I calculated the final score by considering all the data I had, but this was a problem as the data which we have is interdependent on each other, as the emotion scores we have are extracted from the video of the candidate and includes emotions like happiness, anger etc. Whereas the transcript data according to me is extracted from the tone and words used by the candidate during the video, i.e. from applying NLP models on the transcript of each candidate. Hence, this data should have codependency.

There was another problem that the candidates total score was calculated based on the overall average of the data provided. This needed to be analyzed further that whether or not Average was the right overall measure for the data.

1. Consistency Analysis

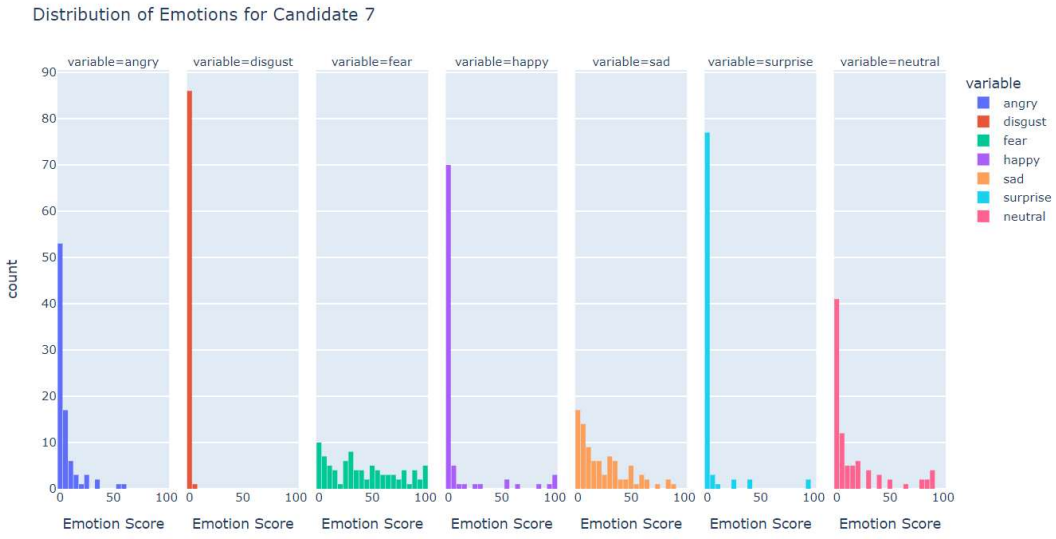
Plotted raw emotion data with standard deviation and moving averages to assess emotional consistency over time.
Analyzed emotional transitions and variability.

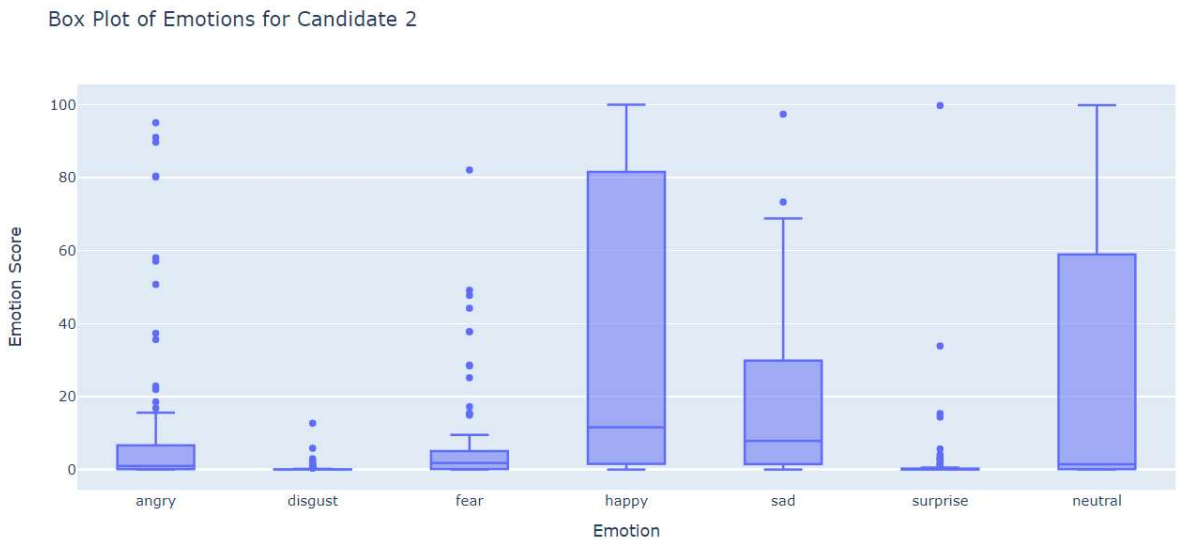




Constructed similar graphs for every candidate but they weren't very useful in our case as moving average is more useful in the case where we have a lot of data.

Created histograms and box plots to identify potential outliers in emotion data.
Examined the distribution of emotions for each candidate.

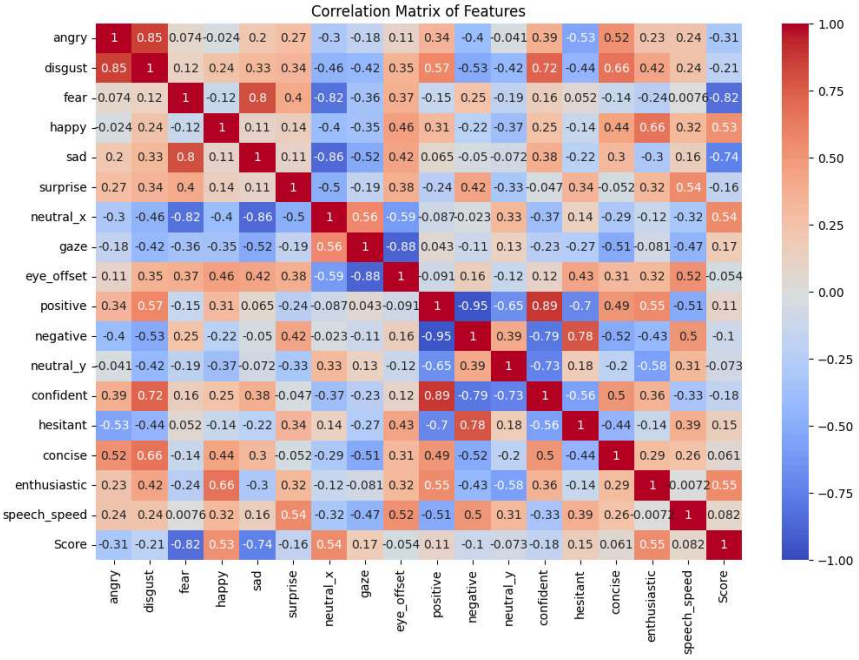




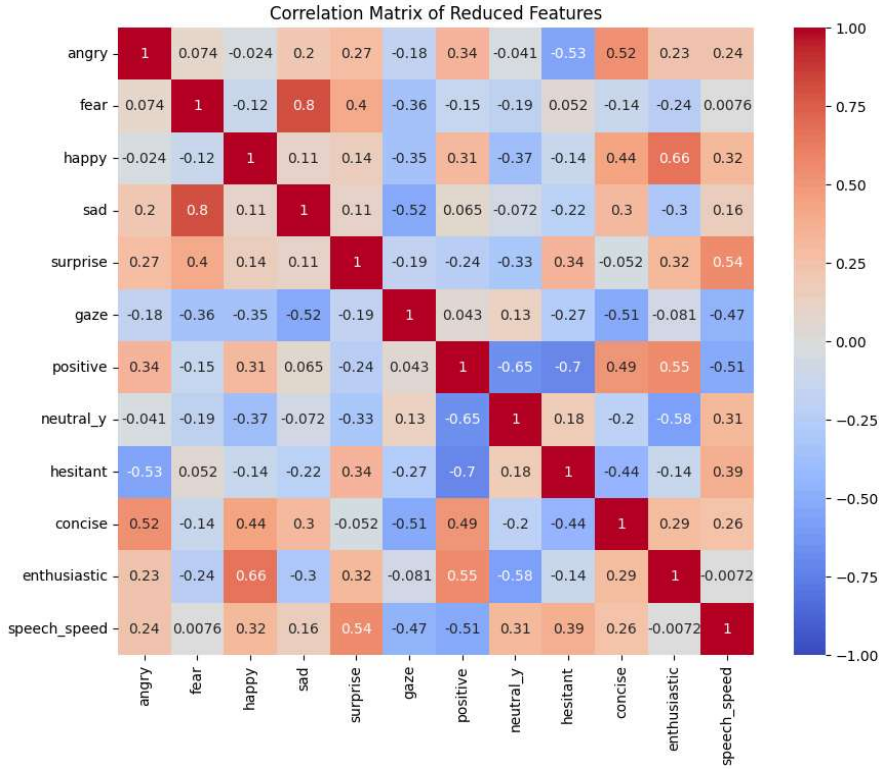
By looking at these plots I understood that we do have outliers and Average may not be the best method for the analysis of our data but at the same time there were some candidates which didn't have significant outliers for certain emotions.
Hence I considered Average to be the best measure.

2. Ruling out some Features and on the basis of that Results

To solve the codependency problem, I constructed Head maps to look at the codependencies.



As predicted this data had a lot of features which were codependent. Removed the features whose codependency is more than 0.8.



As seen above the codependencies were significantly reduced and now I calculated the communication skills and added weights to these features and on the basis of this score suggestions were made that which candidate was suited for the job (Assumed that the jobs for which the selection was being done are(Software Engineer, Data Analyst, UI/UX Designer, Business Analyst)

	candidate_id	communication_skills	recommended_role
0	1	Low	NA
1	2	High	Software Engineer
2	3	Low	NA
3	4	Low	NA
4	5	Low	NA
5	6	High	Software Engineer
6	7	Low	NA
7	8	Low	NA
8	9	Medium	Data Scientist
9	10	Low	NA

4. Skillset and Other Data Analysis

Created a DataFrame containing additional information about candidates, including education, experience, skillset, achievements, extracurricular activities, hobbies/interests, and career goals.
Used this data to suggest suitable roles for each candidate based on their background and skills.

	Serial No.	Education	Experience	Skillset	Achievements	Extracurricular Activities	Hobbies/Interests	Career Goals	Suitable Roles
0	1	B.Tech in Biotechnology, M.Tech from IIT Khara...	Medical Writer at Ciro Klein Farm, Experience ...	Expertise in Regulatory Affairs, Attention to ...	Best Research Award from IIT Kharagpur	Baking, Traveling	Baking, Traveling	Short Term: Apply skills, Long Term: Value Cre...	Regulatory Affairs Specialist, Research Analyst
1	2	BBA	Interned at boutique investment bank, Interned...	Finance, Business Model Preparation	Developed Venture Network Framework	None mentioned	None mentioned	Seeking challenging role in Business Developme...	Finance Manager, Business Development Manager
2	3	B.Tech in Engineering, MBA	Interned as Sales Associate, Interned in Accou...	Sales, Accounting, Content Creation	Guitar Player, YouTube Channel	Music, Traveling, Learning	Music, Traveling, Learning	Seeking challenging and rewarding role	Sales Manager, Financial Analyst, Content Creator
3	4	Engineering Graduate in Electronics and Commun...	Interned at PSK VLSI Design Center, Worked as ...	Data Science, Consulting, Event Coordination	Event Coordinator, Student Leader, Sun NGO Member	Adaptive Learning, Exploring	Adaptive Learning, Exploring	Utilizing knowledge and experience to help com...	Data Scientist, Consultant, Event Coordinator
4	5	Undergraduation in Mass Media, Certifications ...	Won International Art Competition, Experience ...	Art, Writing, Creativity	International Art Competition Winner	Drawing, Painting, Singing, Instagram Reviews	Art, Mental Health Awareness	Applying AI to aid neurodevelopmental disorders	Art Director, Mental Health Advocate, Social M...
5	6	First Year MBA (IIM Kashipur)	3 years at Deloitte (Consulting), Media and PR...	Analytics, Consulting, Strategy, PR Management	Consulting Experience at Deloitte	Media and PR Activities	Learning, Analytics	Interest in Analytics and Mental Health	Analytics Consultant, PR Manager, Content Stra...
6	7	Undergraduate in Earth Science	Worked at GIC Re (Retrocession and Reinsurance...	Reinsurance, Market Understanding, Communication	Experience at GIC Re	None mentioned	Exploring new experiences	Interest in AI-driven social impact	Reinsurance Analyst, Market Analyst
7	8	PGP Finance (IIM Co-Ecode), CA, CFA Level 1	Interned at PWC (Statutory Audit), Worked at I...	Analytical Skills, Financial Analysis, Auditing	CA and CFA Level 1 Cleared	None mentioned	Applying skills in EdTech and creating value	Applying skills in EdTech and creating value	Financial Auditor, Internal Auditor, Finance C...
8	9	B.Tech in Agriculture Engineering, M.Tech in F...	Co-founded Agritech Startup, Project on Remote...	Agritech, Remote Sensing, IoT, AI, Entrepreneu...	Agritech Startup Co-founder	None mentioned	Working in AI-driven challenging areas	Working in AI-driven challenging areas	Agritech Specialist, AI Project Manager, Busin...
9	10	B.Com Honours	Interned as Accounting Associate and Tax Assoc...	Accounting, Taxation, Leadership	Captain of Student Committee, Best Student Award	Bad Scouts and Guides	Reading, Creative Exploration	Short Term: Apply skills, Long Term: Value Cre...	Accounting Specialist, Tax Consultant

For this I trained a lot of Natural Language processing models but they were not anywhere near to the Chat Gpt Model from OpenAI

5. Multiple Approaches for Job Suggestions

In between the primary and secondary analysis I had multiple ideas which I worked on. One of the was to create a list of some **Exhaustive Skills**, the idea was to list a set of skills which are needed for every job in any field, and the basis of these skills could be the data which we have and we could create new features which were nothing but functions of the base features which we already had, similar thing was observed in the Transcript data that the total of Positive, Negative and Neutral was found to be one and the confidence data was calculated by keeping these data as basis. Hence I applied some prompt engineering skills and some intuitions to find the same. The summary of what was did is as follows:

Soft Skills Calculation

Developed a method to calculate soft skills (e.g., confidence, enthusiasm, emotional intelligence) based on emotion and behavioral scores.
Normalized the soft skills scores using min-max scaling.

Job Suggestion Algorithms

Created multiple versions of job suggestion algorithms:

1. Based on raw soft skill scores
2. Based on normalized soft skill scores
3. Based on categorized soft skill levels (Low, Medium, High)
4. A multi-role suggestion algorithm allowing for multiple job recommendations per candidate

Final Scoring Method

Developed a comprehensive scoring method considering various aspects of candidate performance.
Calculated individual scores for different soft skills and an overall composite score.
Provided job suggestions based on the overall score.

6. Key Findings and Insights

1. Emotional Consistency: The analysis revealed varying levels of emotional consistency among candidates. Those with more stable emotional patterns might be better suited for roles requiring composure under pressure.

2. Communication Skills: Transcript analysis provided insights into candidates' communication styles, including their positivity, confidence, and speech patterns. This information is crucial for roles requiring strong interpersonal skills.

3. Gaze Patterns: Analysis of gaze data offered insights into candidates' attentiveness and potential nervousness during the interview process.

4. Soft Skills Profile: The derived soft skills profiles provided a nuanced view of each candidate's strengths, helping to match them with suitable roles.

5. Multiple Job Suggestions: The various job suggestion algorithms demonstrated that candidates might be suitable for multiple roles, highlighting their versatility.

6. Comprehensive Scoring: The final scoring method provided a holistic view of each candidate, considering multiple factors to suggest appropriate job roles.

Conclusion

The analysis provided a detailed evaluation of each candidate’s emotional consistency, communication skills, and overall performance, offering key insights into their suitability for various roles. Emotion data highlighted candidates with stable emotional patterns, ideal for high-pressure environments, while transcript analysis revealed valuable traits like confidence and clarity in communication, essential for roles requiring strong interpersonal skills. Gaze data also provided insights into candidates' attentiveness and potential nervousness during interviews, further contributing to their overall assessment.

Based on the comprehensive scoring system and job suggestion algorithms, Candidates 6, 2, and 8 stood out as strong fits for technical roles. Their emotional consistency, strong communication, and soft skills profiles suggest they would excel in positions like Data Scientist and Software Engineer. This highlights their ability to handle the analytical and problem-solving demands of such roles while maintaining effective interpersonal and technical skills.