Exploratory Data Analysis

This report summarizes my findings after performing exploratory data analysis on the dataset provided.

Dataset Description

The dataset contains emotion scores, transcript scores and corresponding transcripts extracted from the introduction videos of 10 candidates applying for a job. The dataset was divided into three major categories.

1. **Emotion scores**

   The dataset contains the emotions of candidates throughout the video including emotion scores of angry, happy, sad, disgust, fear, surprise and neutral. It also contains binary gaze data, eye offset, blink data and the metadata of the image frames extracted from the video.

2. **Transcript scores**

   The dataset contains positive, negative, neutral, confident, hesitant, concise, enthusiastic scores obtained after sentiment analysis on the transcripts of each candidate.

3. **Transcript text**

   Contains the actual text of the transcript of the videos of each candidate.

Objectives

- Perform prompt engineering and Exploratory Data Analysis on the given dataset to generate valuable and actionable insights from the data.
- Analysis of communication skills and finding areas of expertise based on data.
- Gain any other insights that could be useful in making a decision on the recruitment of the candidate.
- Decide whether the candidate should be recruited or not after a thorough analysis of the data.

Exploratory Data Analysis

Exploratory data analysis was performed on *emotion scores* and *transcript scores* datasets to investigate these datasets and obtain valuable insights about each candidate applying for the job. These are the steps in which the analysis was done.

1. Understanding the given data
2. Analyzation and Visualization of the data
3. Summarize the findings from the data analysis.
4. Decide whether the candidate should be recruited or not after a keeping into consideration both the prompt engineering findings and exploratory data analysis findings.

Data analysis of *transcript text* dataset was done using prompt engineering on ChatGPT to gain information about the candidates from the transcripts.

Understanding the data

1. Emotion scores

These are the percentages of the predicted emotions exhibited by the candidate in each frame of the video. There are seven different emotions taken into consideration: angry, happy, sad, disgust, fear, surprise and neutral. Each of these emotion scores signify something in the context of recruitment.

- Happy: Higher happy scores could imply a candidate with a positive outlook, likely to contribute to a pleasant work atmosphere, team morale, and client interactions.

- Angry or sad: Elevated sad or angry scores might raise concerns about the candidate's emotional resilience, difficulty in handling stress or conflicts, and ability to cope with challenges, which could impact productivity and well-being in a professional environment.

- Fear or disgust: Higher fear or disgust scores could indicate a candidate's potential apprehension or anxiety, or a candidate's strong negative reaction, which might affect their ability to perform under pressure or handle client interactions.

- Surprise: Higher surprise scores might suggest a candidate who is adaptable, open to change, and willing to take on new challenges, which can be beneficial for growth and innovation within the organization.

- Neutral: Balanced neutral scores indicate an emotionally composed candidate, capable of maintaining stability and focus, even in potentially stressful situations.

Apart from these scores, each frame in the video has a corresponding gaze, blink and eye offset values.

- Gaze: Value is either 0 or 1. If gaze is 1, then candidate is looking into the camera. If gaze is 0, then candidate is not looking into the camera.

- Blink: Tells us if candidate is blinking in the frame or not. Value is either 0 or 1.

- Eye offset: Tells us the deviation of the eye from the camera.

2. Transcript Scores:

The transcript is broken down into phrases told by the candidate. For each phrase, there is a positive score, negative score, scores for neutral, confident, hesitant, concise, and enthusiastic determined by sentiment analysis. Each score is between 0 and 1. There is speech speed also for each phrase. Each of these scores signify something in context of recruitment.

- Positive, negative, and neutral scores: A candidate with a higher positive score and lower negative score may convey a more optimistic and positive outlook. A balance of positive and neutral scores could indicate a rational communicator.

- Confident and hesitant scores: A higher confident score signifies assertiveness and conviction in their statements, which is generally valued in roles requiring leadership

or client-facing responsibilities. A high hesitant score might indicate nervousness or lack of confidence, which could be a consideration for roles involving presentations or interactions with stakeholders.

- Concise score: Candidates with higher concise scores tend to communicate efficiently. Being concise is especially important in roles where clear and brief communication is key.

- Enthusiastic score: Candidates with higher enthusiastic scores might come across as more engaged and energetic, which is beneficial for roles that require a high level of passion and commitment.

- Speech speed: A balanced speech speed (2.5 – 3 words per second), not too fast or too slow, can indicate effective communication. Extremely fast speech may be a sign of nervousness or haste, while slow speech might suggest cautiousness or overthinking.

Analyzing and Visualizing the data

1. **Emotion scores:**

   - The first and **most important observation** here was that candidates **5** and **6** do not have enough data for obtaining useful information. Due to the small datasets, findings on emotion scores on these two candidates do not have significant meaning. Therefore, going further in the report, the analysis from emotion scores for these two candidates will not be taken into consideration.

   - The *blink* data is subjective to each candidate and does not provide any meaningful information or relationship with the emotion scores. This data has no significance in the context of recruitment and so is not considered in any analysis.

   - The columns *'gaze'* and *'eye_offset'* from *gaze.csv* of each candidate is merged with *emotion.csv* on each candidate on the common column *'image_seq'*. Now we obtain the **correlation matrix** between emotion scores, gaze and eye offset of each candidate.

   - An important observation from the correlation matrix is that for all the candidates, **gaze** and **eye offset** are **highly negatively correlated**. This intuitively makes sense, as if the candidate looks into the camera for a longer time, eye offset is going to be low and if the candidate looks into the camera for a short time, eye offset is going to be high. Since these columns are highly correlated, we consider only gaze and not eye_offset during further analysis.

   - The **mean emotion scores** are obtained for each candidate and displayed in **pie charts**. The dominant emotion for each candidate is noted.

   - **Step plots** are obtained for each candidate using **gaze** data. The percentage of time the candidate looks into camera is determined.

2. **Transcript scores:**

   - The important columns from the transcript data csv files of each candidate are the columns of each of the transcript scores and the *speech_speed* of the corresponding text in the *text* column.

   - The mean transcript scores are plotted in a bar graph for each candidate. These bar graphs help in visualizing the sentiment of the transcripts for each of the candidates.

   - Line plots of speech speed is obtained for each candidate.

The graphs obtained upon analysis are shown in the next part of the report. The next part of the report contains findings for each candidate individually.

Findings for each candidate

Until now, we have used prompt engineering to get to know about the candidates from the transcripts (this information is stored in *candidates_info.csv*) and we have performed data analysis on the emotion scores and transcript scores. Now we summarize the findings from prompt engineering and data analysis for each of the candidate and make a decision on whether we can recruit the candidate or not.

Candidate 1

Name: Jeffrey Shepherd

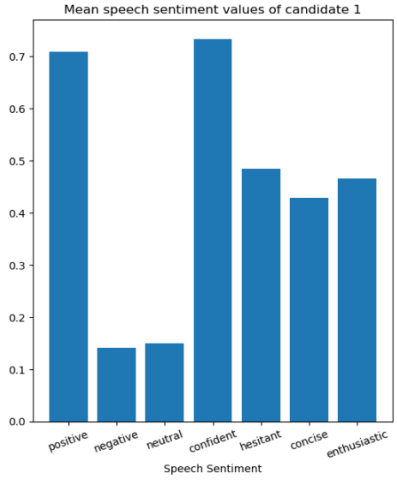
Education: Postgraduate and management from IIM Coikode. B.Tech in Biotechnology from Heritage Institute of Technology Kolkata. M.Tech from IIT Kharagpur.

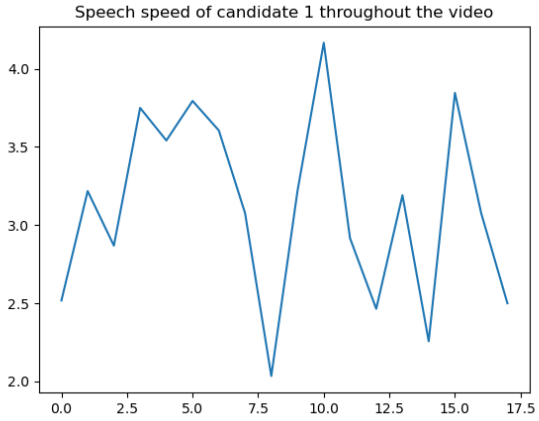
Work Experience: Three years in the regulatory affairs domain of the pharmaceutical industry. Worked as a medical writer in Ciro Klein Farm, Mumbai specialized in drug safety and risk management.

Relevant Skills: Regulatory affairs. Medical writing. Drug safety. Risk management

Other details: Exhibits attention to detail, consistency in academics, and an analytical mindset. Demonstrates a proactive approach to problem-solving and viewing challenges from multiple angles.

Language Proficiency: Good

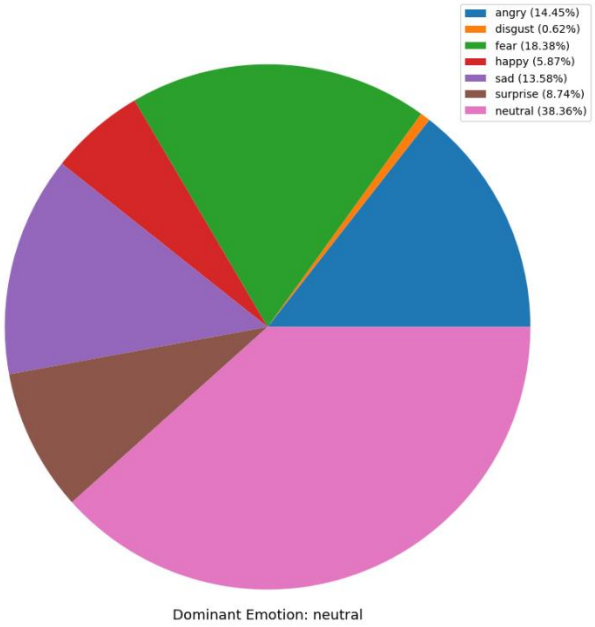




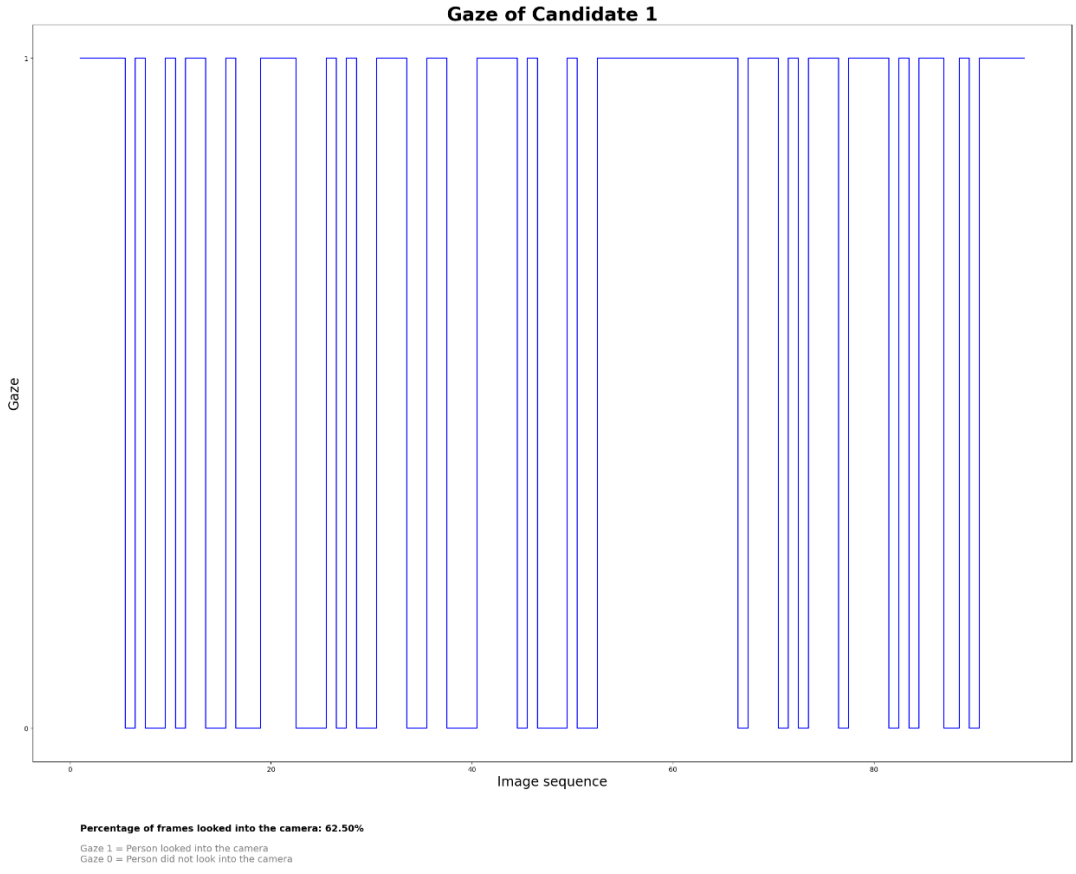
The above plots show the mean values of the transcript speech sentiment scores and the variation in speech speed throughout the video.

The plot on the left shows that the candidate is **highly positive and confident** through his words.

The speech speed plot shows that the candidate speaks at a pace of 3 – 3.5 words per minute. Although there are some fluctuations in the speed, this might be due to pauses taken between few sentences.



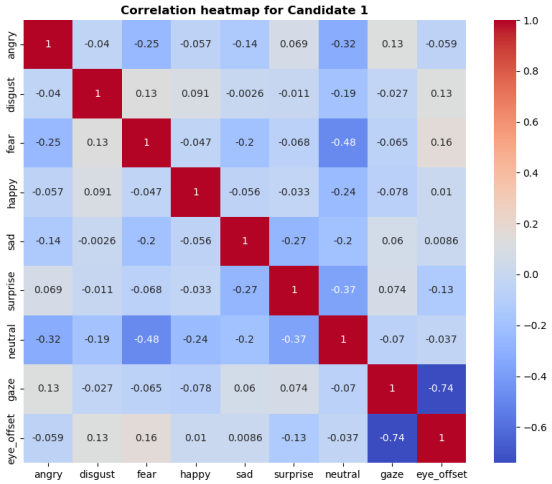
Emotions Distribution for Candidate 1



Percentage of frames looked into the camera: 62.50%

Gaze 1 = Person looked into the camera
Gaze 0 = Person did not look into the camera

The emotion distribution plot shows that the candidate is **fairly neutral** indicating composure and stability. He is seeing the camera for satisfactory part of the duration of the video.



The correlation heatmap does not provide any new information other than the already analyzed high negative correlation between eye offset and gaze.

Keeping in mind the strong background in biotechnology and experience in the pharmaceutical industry of the candidate and the positivity and confidence in his thoughts and words, he could be an asset in Tech or Research & Development departments in a company. Therefore, this candidate **must be recruited**.

Candidate 2

Name: Cameron Barajas

Education: BBA in 2022

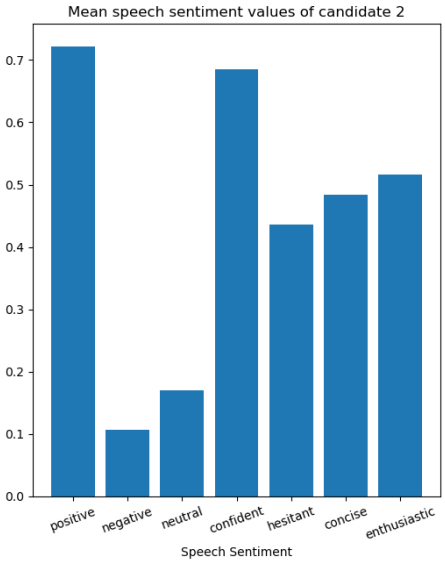
Work Experience: Internship experience in a boutique investment bank and a startup (Kabadi Techno) focusing on finance.

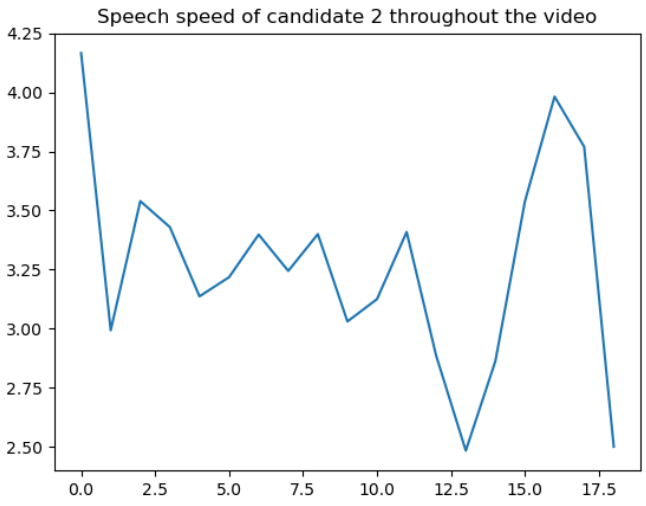
Relevant Skills: Finance. Venture network. Financial modeling

Other details: Expresses a passion for challenges and a desire to grow at an accelerated rate within a challenging role.

Language Proficiency: Satisfactory

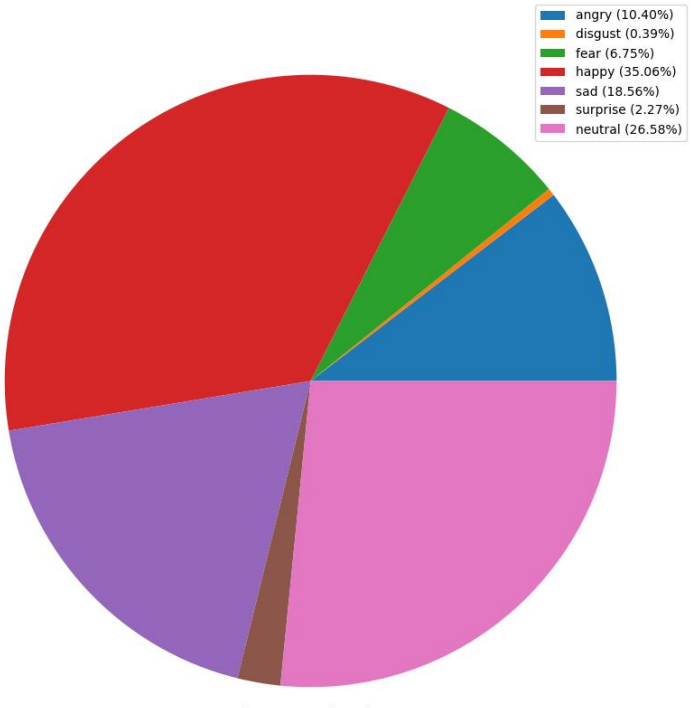
The below plots show the mean transcript sentiment scores and the speech speed of the candidate.



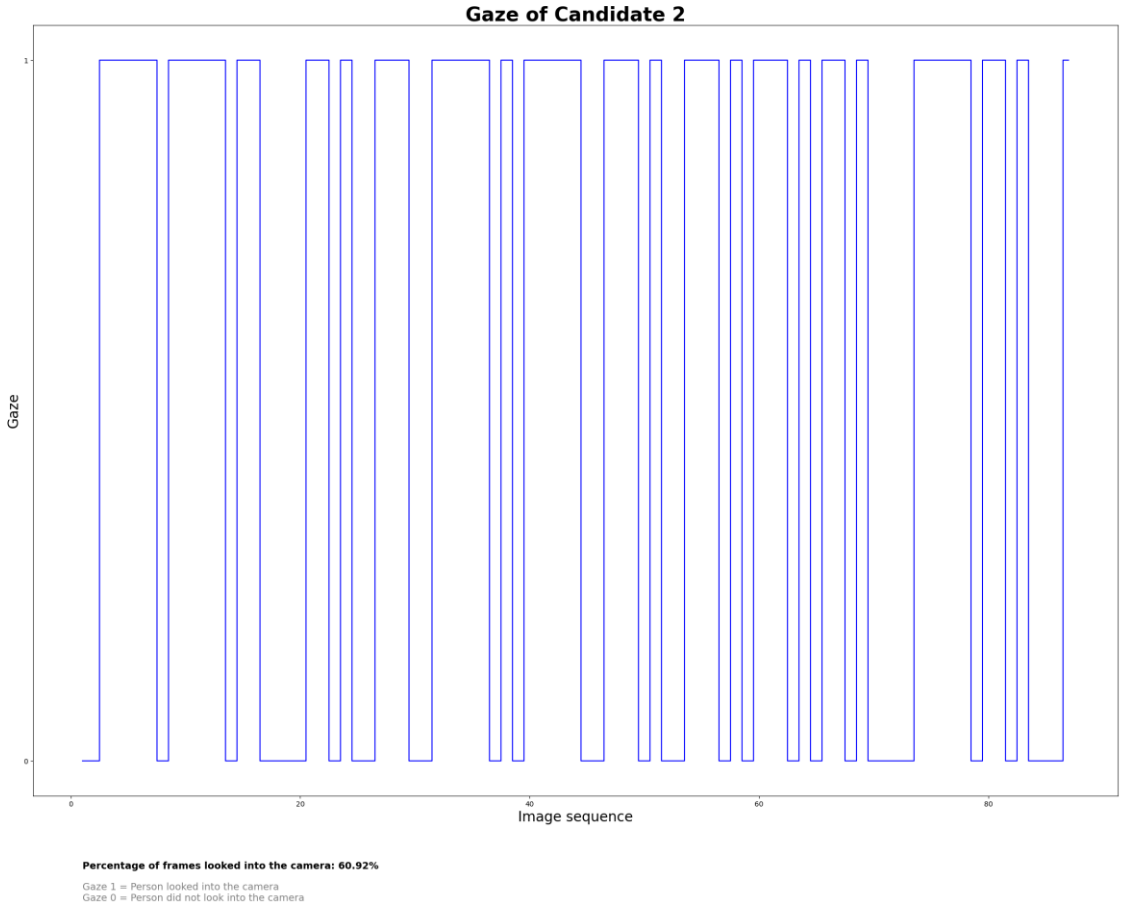


The plot on the left shows that the candidate is **highly positive and confident** through his words.

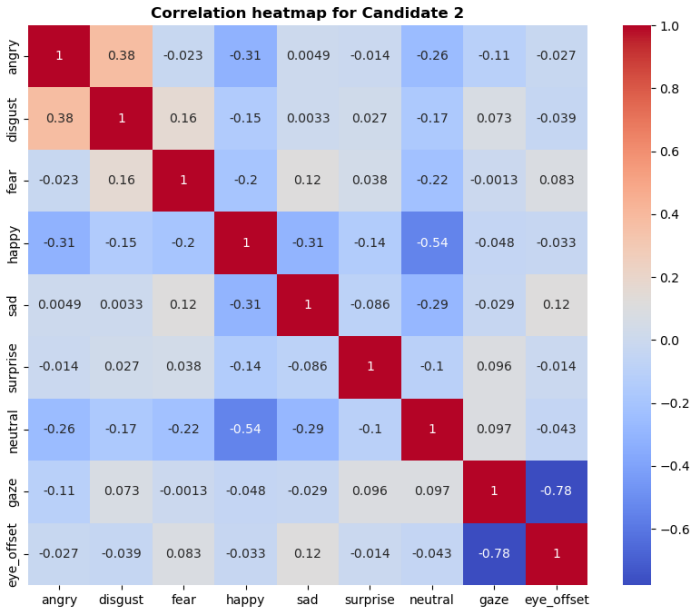
The plot on the right shows that the candidate speaks at a **fairly stable pace** of 3 – 3.5 words per minute.



Emotions Distribution for Candidate 2



The emotion distribution plot shows that the candidate is **happy or neutral** for the majority of the video indicating composure and stability and eagerness. He is seeing the camera for satisfactory part of the duration of the video.



The correlation heatmap shows **negative correlation between happy and neutral** indicating that he is either neutral or happy. This was already deduced from the pie chart.

The candidate's educational background and work experience suggest that the candidate would be suited for finance firms or in finance departments. The candidate is a recent graduate and is a fresher with two internship experiences. With his positive and happy nature, combines with his passion to work would make him a **good fresher to be recruited**.

Candidate 3

Name: Michael Guzman

Education: Undergraduate entrance exam of BHU with top 1.2% rank. Pursued honors from Varanasi University.

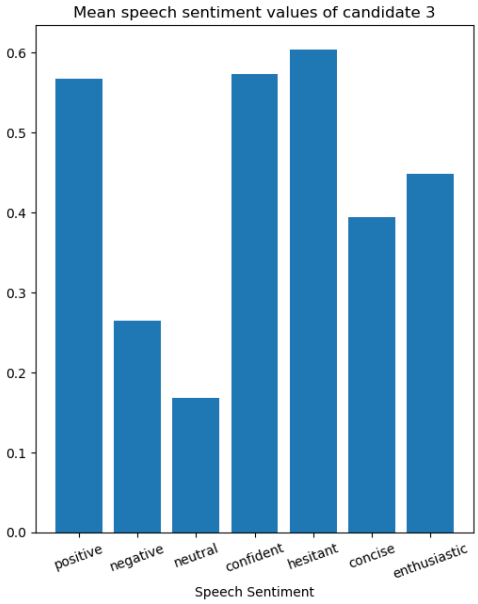
Work Experience: Internship in a steel manufacturing firm and an accounting firm.

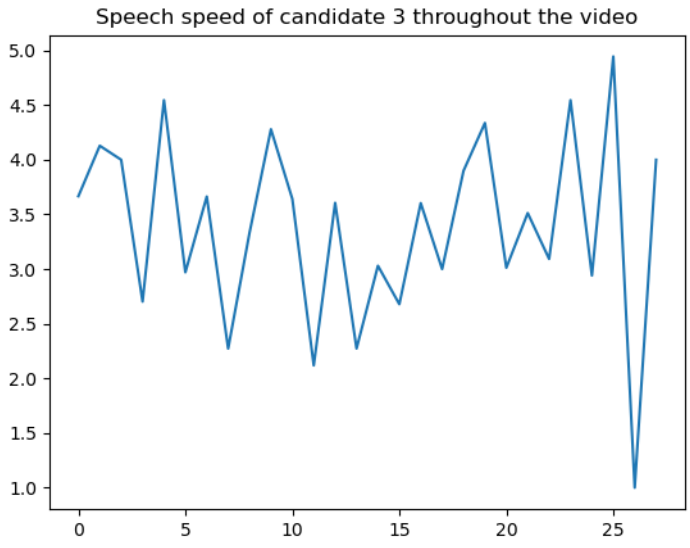
Relevant Skills: Sales. Accounting. Guitar playing. Fingerstyle guitar. YouTube content creation.

Other details: Exhibits a diverse skill set including sales, accounting, and guitar playing, demonstrating adaptability and versatility.

Language Proficiency: Good

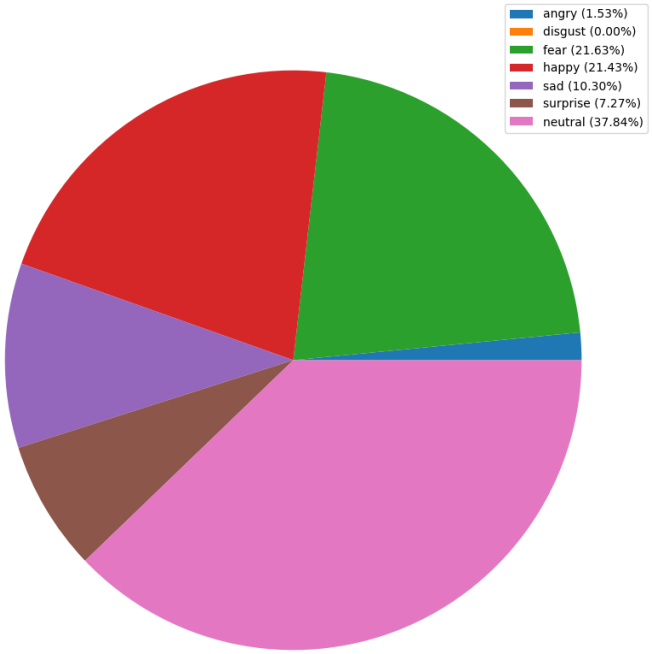
The below plots show the mean transcript sentiment scores and the speech speed of the candidate.





The plot on the left shows that the candidate shows mixed sentiments of **positivity, confidence and hesitation** through his words.

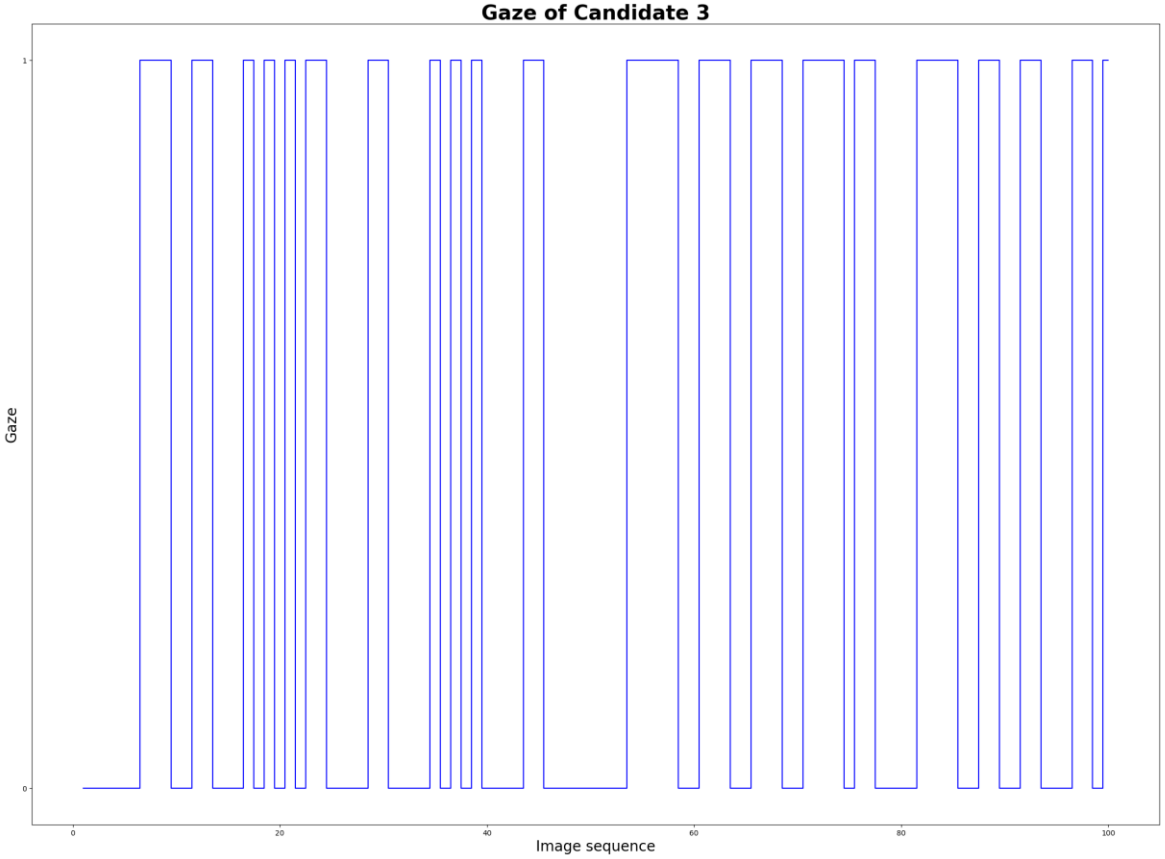
The plot on the right shows that the candidate speaks at a **hurried pace** of 3.5 to 4 words per minute while sometimes speaking at 4.5 words per minute. This might indicate some anxiety or nervousness on the candidate.



Dominant Emotion: neutral

Emotions Distribution for Candidate 3

Gaze of Candidate 3

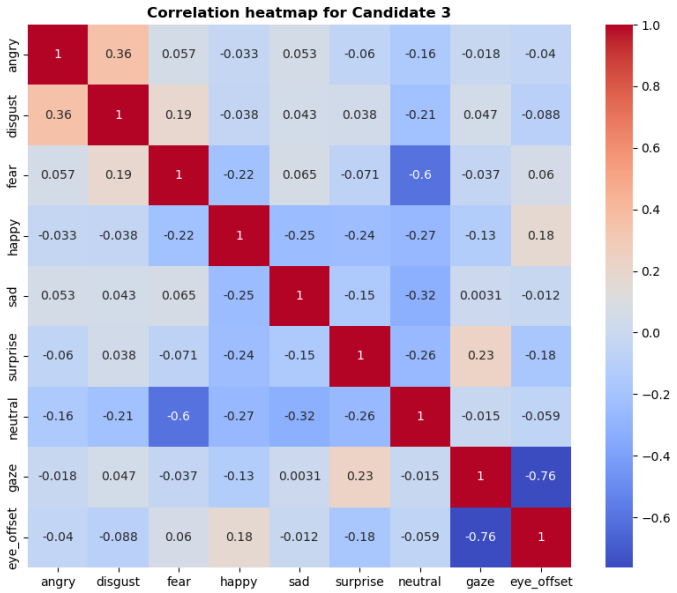


Percentage of frames looked into the camera: 45.00%

Gaze 1 = Person looked into the camera
Gaze 0 = Person did not look into the camera

The pie chart shows that the candidate is **mostly neutral** throughout the video although he shows **fear** in some parts of the video.

The step plot denotes that the candidate has **not seen the camera** for even half the time of the video.



The correlation heatmap shows **negative correlation between fear and neutral** indicating that he is either neutral or scared. This was already deduced from the pie chart.

After the analysis, we can realize that even though the candidate has proven well academically with top scores in undergraduate entrance exam and an undergraduate honors degree, the candidate express anxiety and nervousness in both his words and emotions. Therefore, the candidate **cannot be recruited** as his characteristics could affect the productivity or performance under high pressure which is not suitable in a work environment.

Candidate 4

Name: Monique Mccormick

Education: Engineering graduate in electronics and communication field. GATE rank 5300

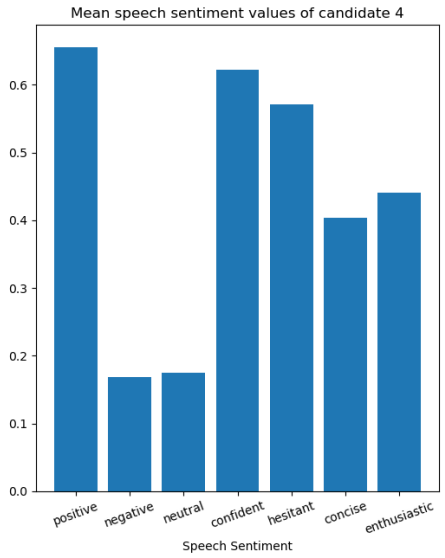
Work Experience: Internship at PSK VLSI Design Center. Worked as an academic advisor in a school for 19 months

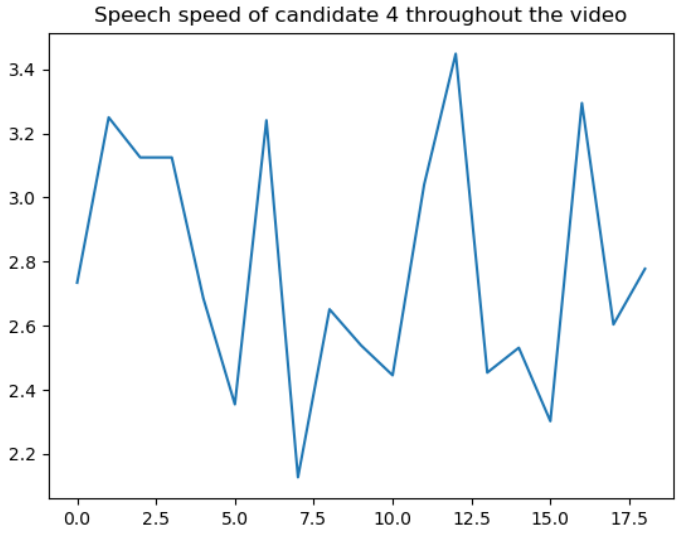
Relevant Skills: VLSI design. Academic advising.

Other details: Enjoys sports like badminton and chess, highlighting a well-rounded personality and an interest in physical activities.

Language Proficiency: Good

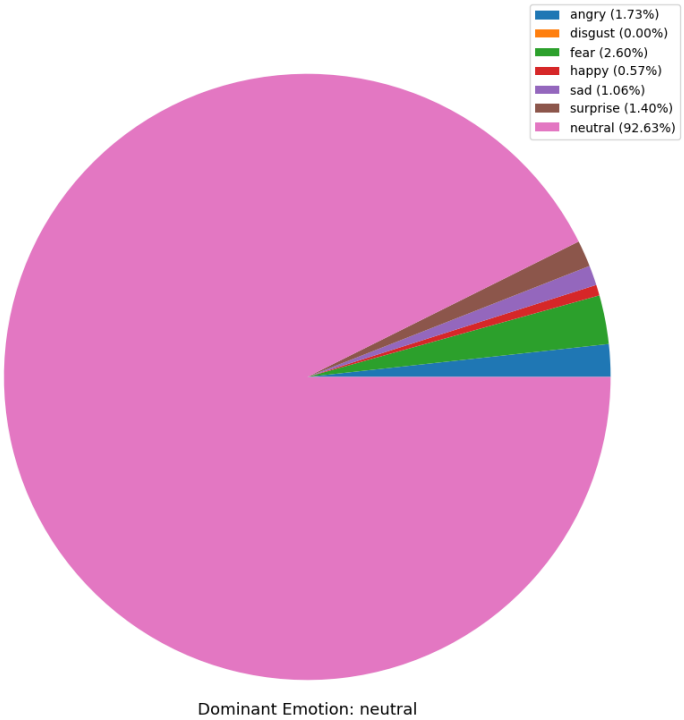
The below plots show the mean transcript sentiment scores and the speech speed of the candidate.





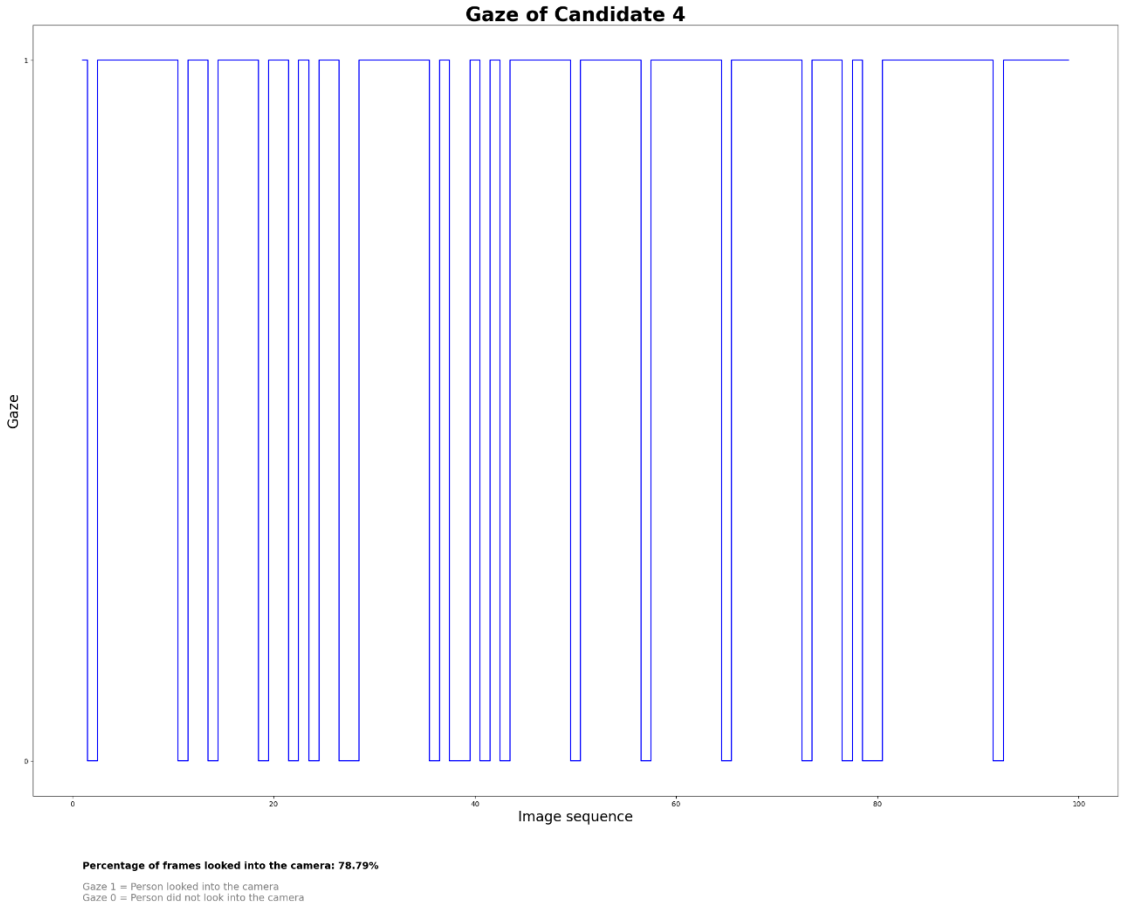
The plot on the left shows that the candidate shows mixed sentiments of **positivity, confidence and hesitation** through their words, although mostly positive.

The plot on the right shows that the candidate speaks at a **stable pace** of 2.4 to 3 words per minute.

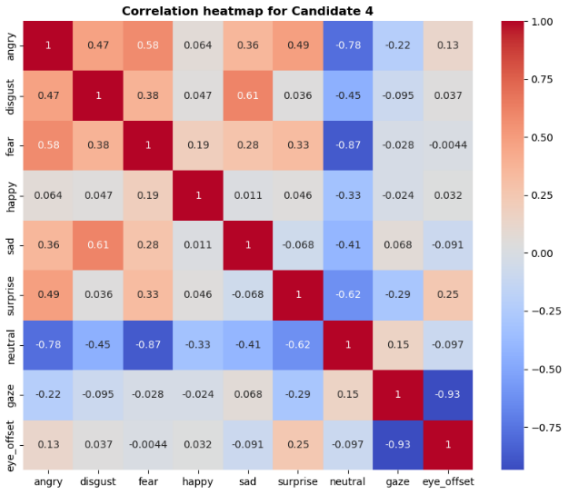


Emotions Distribution for Candidate 4

The pie chart shows that the candidate is **neutral** for the entirety of the video indicating that she is emotionally composed.



The candidate has looked into the camera for majority of the time, indicating confidence in presentation of themselves.



The correlation heatmap shows **negative correlation between neutral and fear,angry.** This indicates that the candidate is either neutral or scared and angry. From the above analysis, the candidate is very neutral and so we can expect them to be less scared or angry.

Overall, the candidate is confident. The candidate also has a great educational background and a good experience in academic advising. Therefore, Monique Mccormick is a **potential candidate** for **management or strategic roles** in a company.

Candidate 5

Name: Sakshi

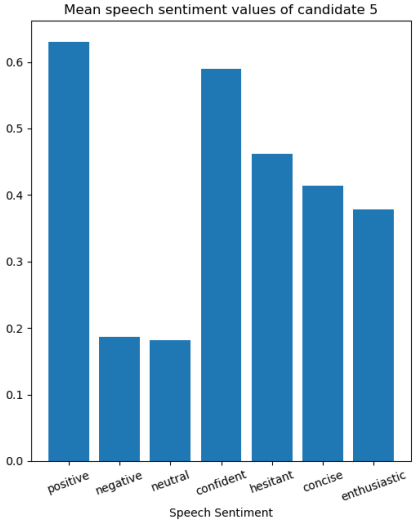
Education: Undergraduate in mass media with specialization in advertising

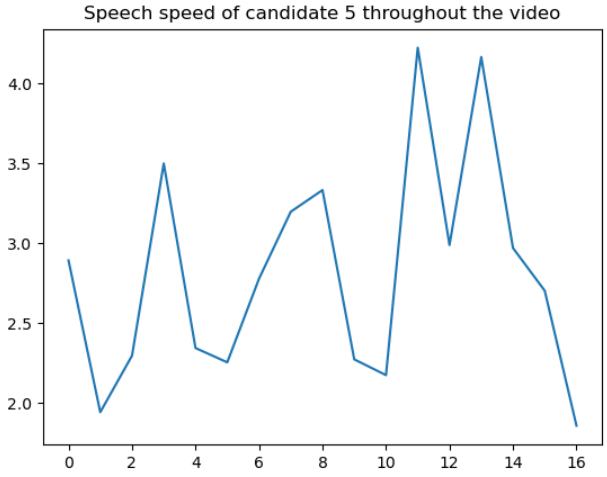
Work Experience: Participation in international art competition hosted by Krezkazad

Relevant Skills: Advertising. Entrepreneurship. Foundations of management

Other details: Passionate about raising awareness for mental health issues, with innovative ideas integrating AI and education.

Language Proficiency: Good

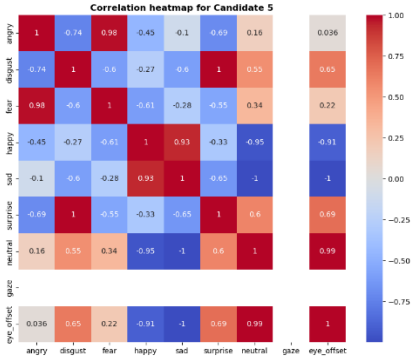


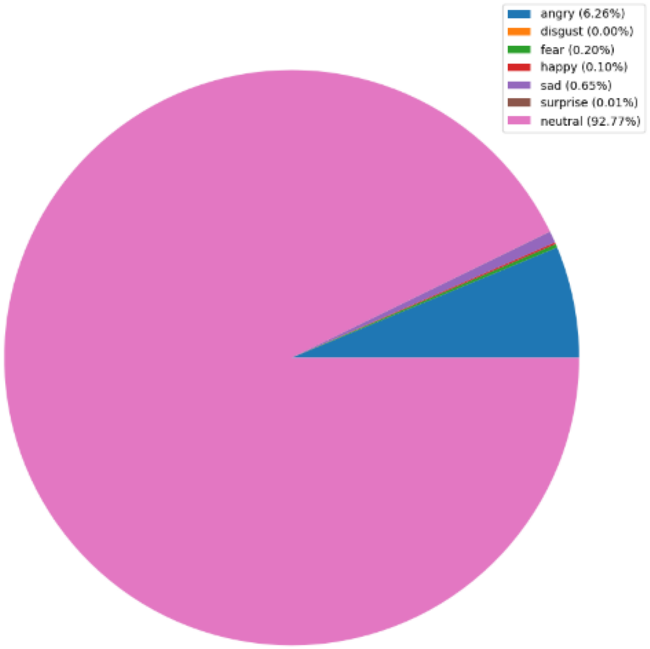


The plot on the left shows that the candidate shows **positivity and confidence** through their words.

The plot on the right shows that the candidate speaks at a **fluctuating pace** of 2 to 4 words per minute indicating less clarity of thoughts.

The below heatmap and plots show the correlation matrix, the emotional distribution and gaze data throughout the video of the candidate. Note that, as seen from the gaze data, there are only 4 image sequences or image frames of video denoting that the dataset is too small to do any analysis on. Therefore no insights are taken from the following plots and correlation heatmap.

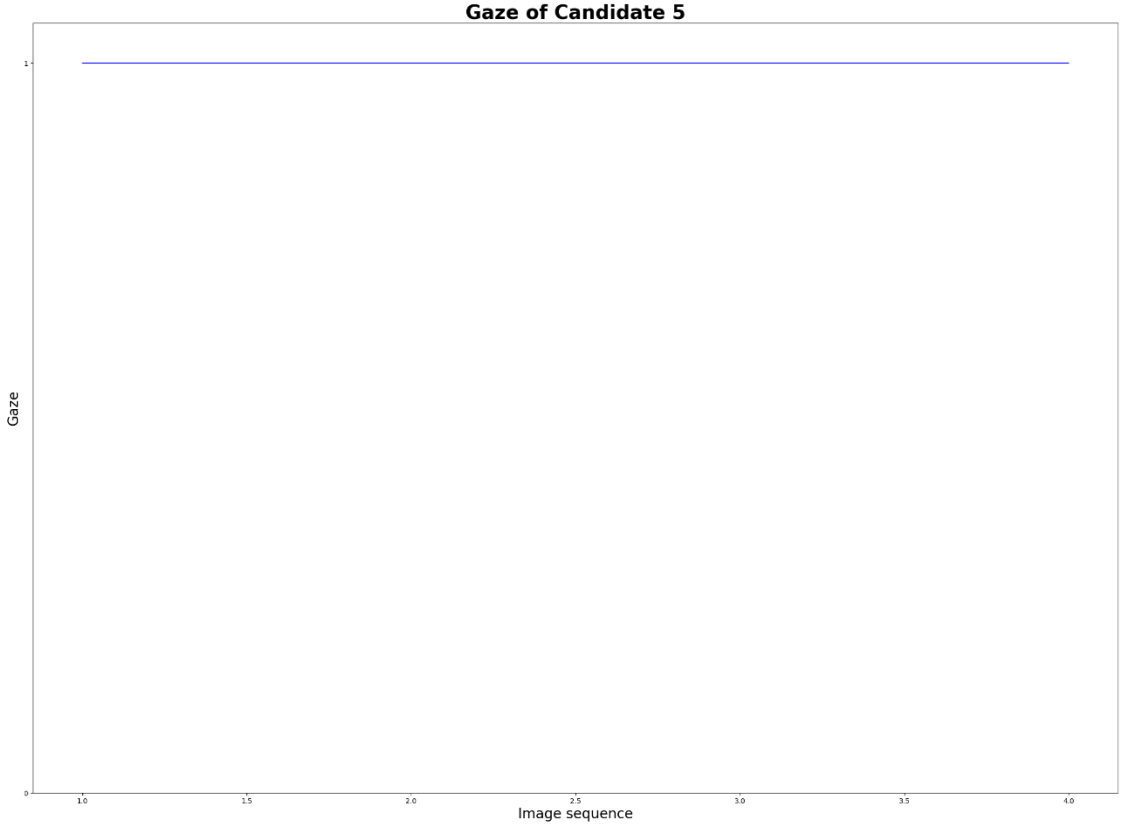




Dominant Emotion: neutral

Emotions Distribution for Candidate 5

Gaze of Candidate 5



We do not have a good emotion dataset to judge the candidate through the video. We do not know if the candidate is a fresher or completed her undergraduation some years ago but she does not have any work experience including internships to her name. Therefore, there is no evidence present to justify that the candidate could work and perform well in a company and so she **cannot be recruited**.

Candidate 6

Name: Nathan Lewis

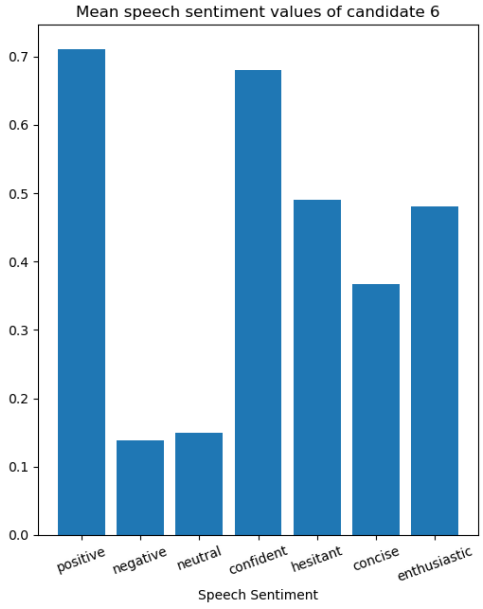
Education: Engineering graduate. Currently pursuing MBA analytics from IIM Kashipur. Three years of consulting experience at Deloitte.

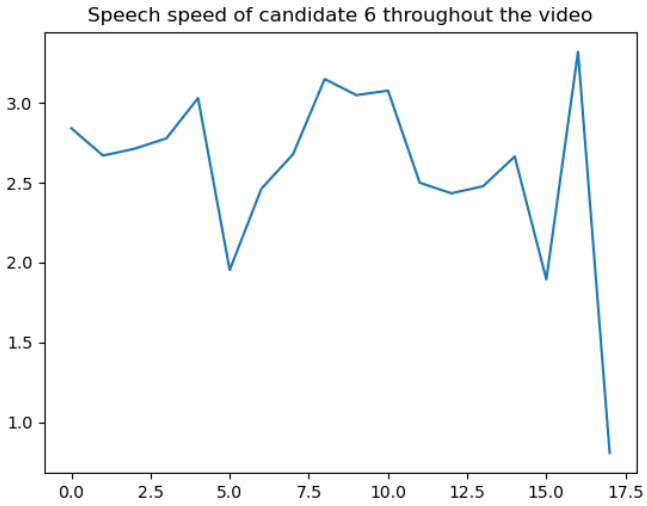
Work Experience: Validation processes for pharmaceutical software. Social media management and content writing for college.

Relevant Skills: Analytics. Strategy planning. Social media management.

Other details: Strong communication skills and a passion for understanding people and their needs.

Language Proficiency: Good

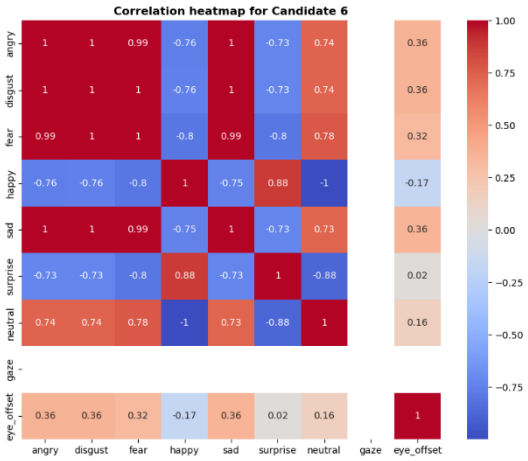


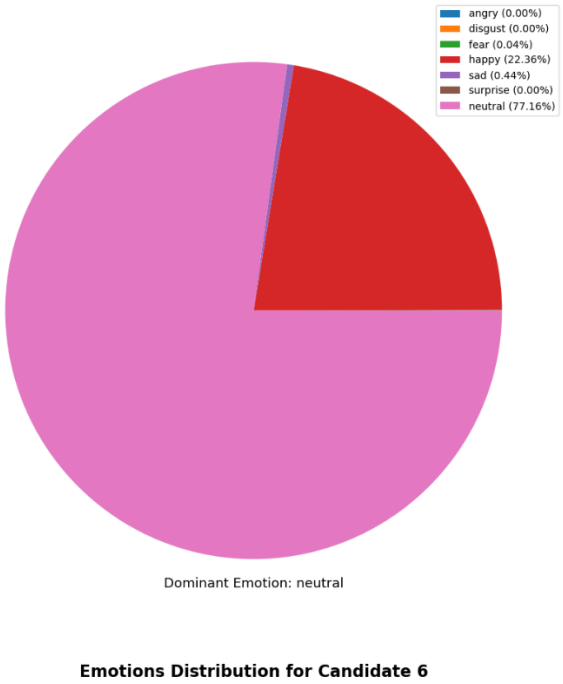


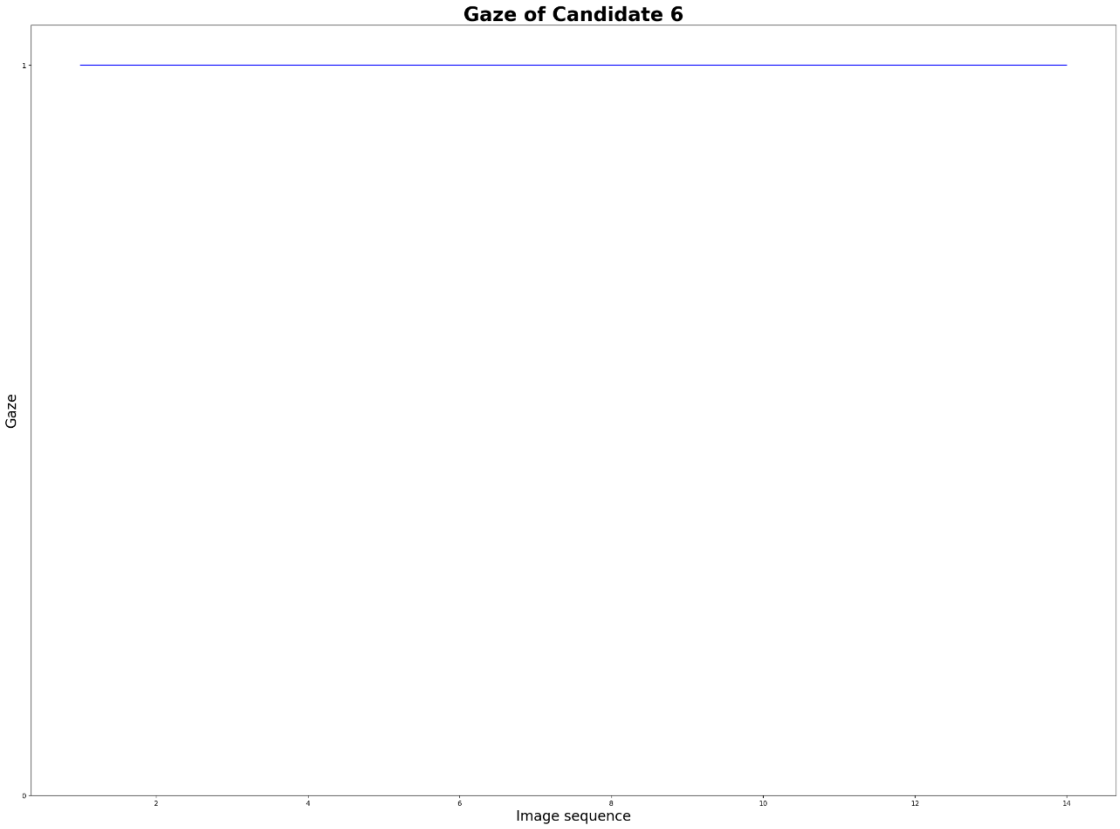
The plot on the left shows that the candidate shows **positivity and confidence** through their words.

The plot on the right shows that the candidate speaks at a **stable pace** of 2.5 to 3 words per minute.

The below heatmap and plots show the correlation matrix, the emotional distribution and gaze data throughout the video of the candidate. Note that, as seen from the gaze data, there are only 14 image sequences or image frames of video denoting that the dataset is too small to do any analysis on. Therefore no insights are taken from the following plots and correlation heatmap.







Although we do not have a good emotion dataset to judge the candidate through the video, the transcript and speed speech shows that the candidate is a positive and confident person. Apart from that, the candidate has a strong academic background with great work experiences. Being an analytics students and experience in consulting, he is an asset to **management and strategic departments** who can aid in analytics-driven decision making. His experience in social media management also adds worth to his already good resume. Overall, he is a great candidate who **must be recruited**.

Candidate 7

Name: Joseph Nichols

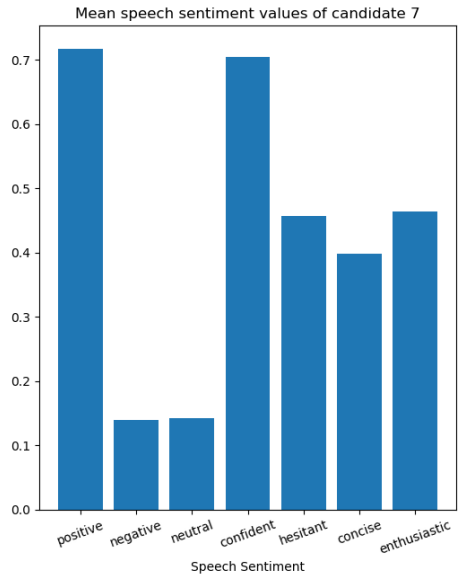
Education: Undergraduate in earth science from Banaras Hindu University

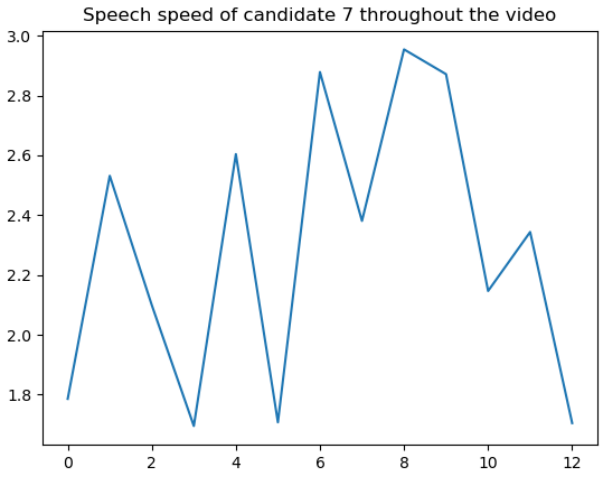
Work Experience: Worked in reinsurance underwriting departments of General Insurance Corporation of India.

Relevant Skills: Reinsurance. Analytical skills.

Other details: Eager for new experiences and values uniqueness in individuals, aligning with the company's ethos.

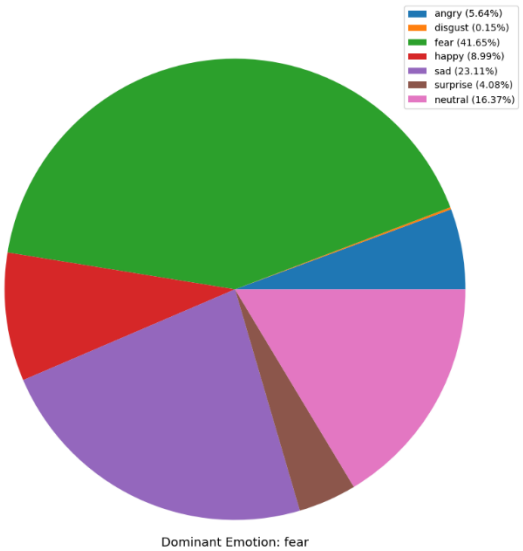
Language Proficiency: Good





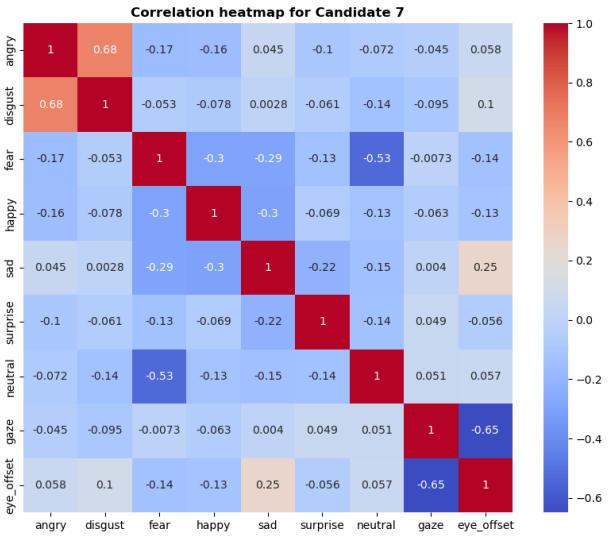
The plot on the left shows that the candidate shows **positivity and confidence** through their words.

The plot on the right shows that a **fluctuating pace** of speech speed. But, even then the speed is neither too slow nor too fast.



Emotions Distribution for Candidate 7

These two plots above show that even though the candidate looks into the camera for a majority of the time, he displays emotions of **fear or sadness**.



The **negative correlation between neutral and fear** denotes that he either displays neutral emotions or is scared. The **positive correlation between angry and disgust** means that he is angry when disgusted and vice versa. But in the major part of the video, he doesn't show any emotions of anger or disgust.

The candidate although shows fear or sadness in the video, otherwise displays positivity and confidence in his words. His experience in underwriting and skills in analytics could be useful in **Human Resource departments**. The candidate **can be recruited** but **should not be the primary choice** in case of vacancy in said roles.

Candidate 8

Name: Srivats Biyani

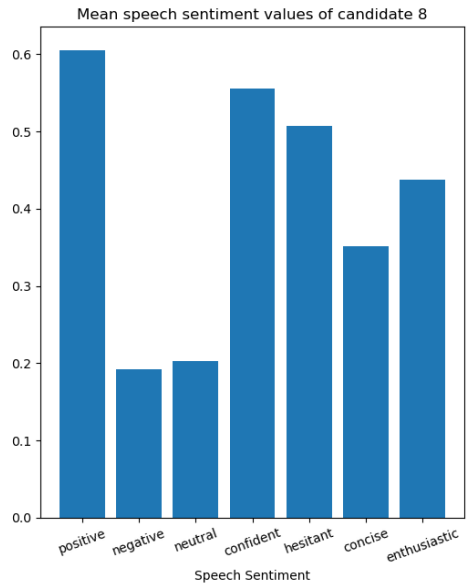
Education: Chartered accountant. Cleared CFA level 1. PGP finance student at IIM Co-Ecode.

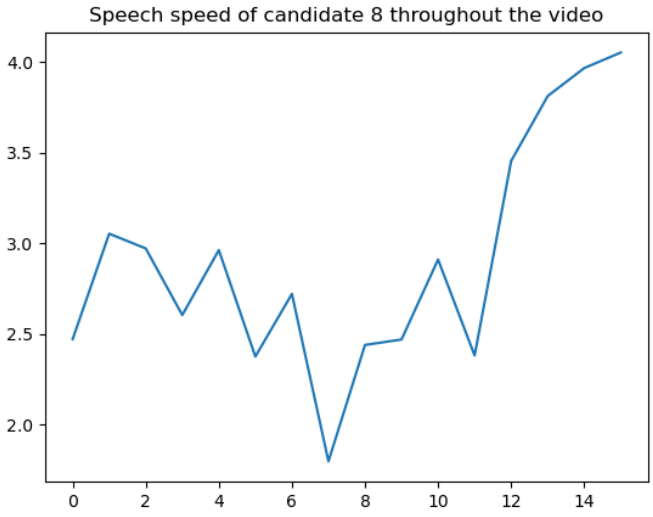
Work Experience: Internship with PWC. Worked in internal audit at ITC limited.

Relevant Skills: Accounting. Finance. Analytics.

Other details: Passionate about analytics and education. Aims to expand successful edtech models.

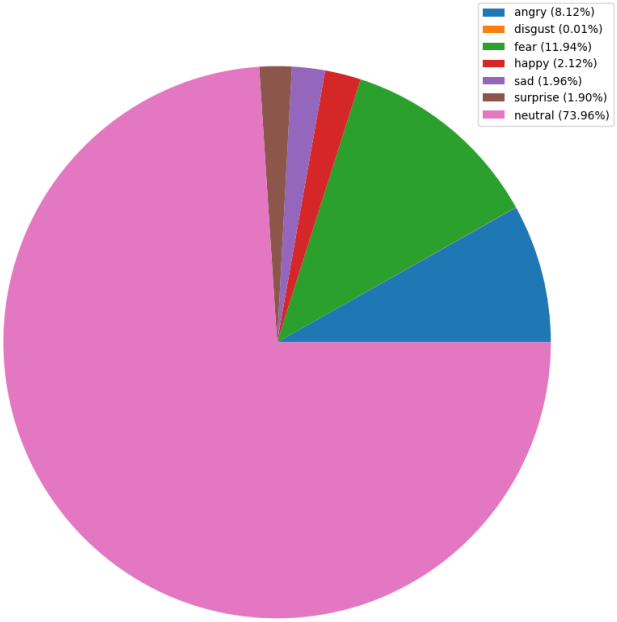
Language Proficiency: Good





The plot on the left shows that the candidate shows **positivity and confidence** through their words.

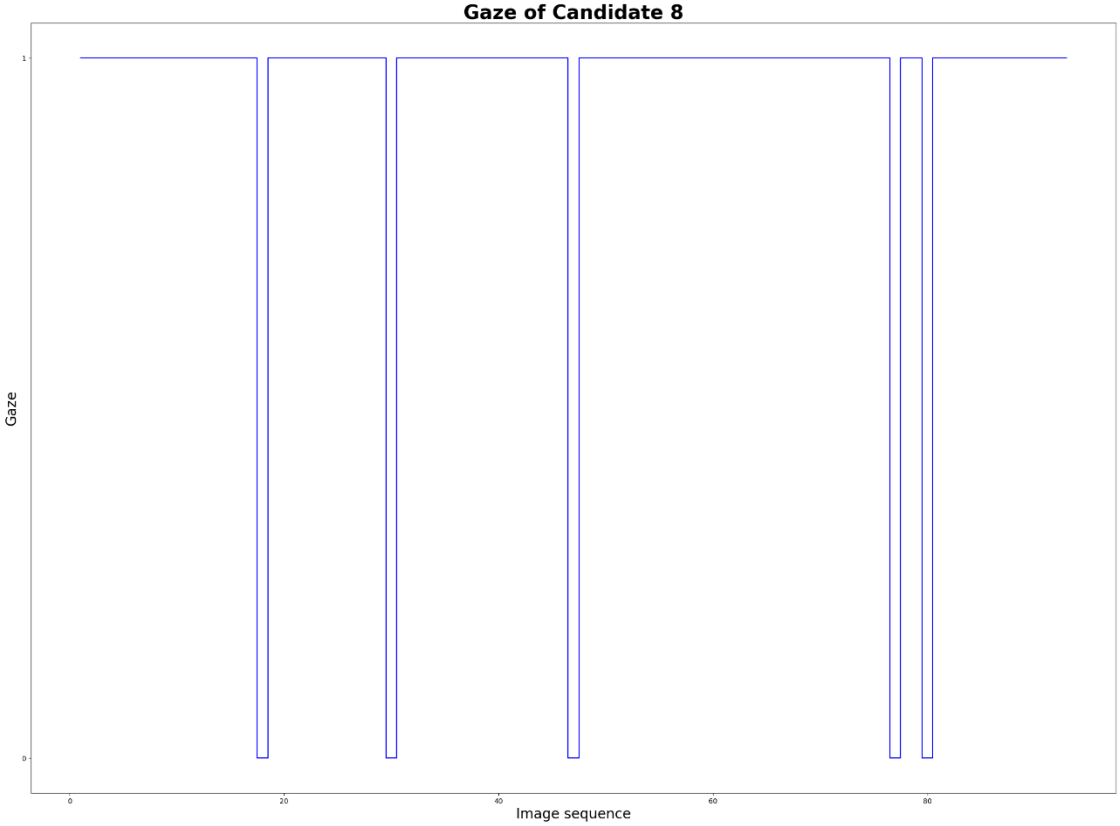
The plot on the right shows that a candidate is speaking at a **stable pace** except for the last part of the video.



Dominant Emotion: neutral

Emotions Distribution for Candidate 8

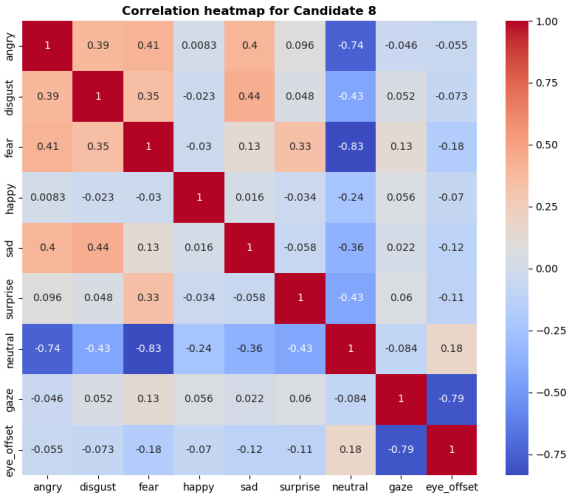
Gaze of Candidate 8



Percentage of frames looked into the camera: 94.62%

Gaze 1 = Person looked into the camera
Gaze 0 = Person did not look into the camera

The above two plots show that the candidate expresses neutral emotions and is looking in the camera for the entirety of the video. This indicates that the candidate is confident and composed.



The correlation heatmap shows **negative correlation between neutral and fear, angry.** This indicates that the candidate is either neutral or scared and angry. From the above analysis, the candidate is very neutral and so we can expect him to be less scared or angry.

Overall the candidate is very positive and confident in his words and emotions displayed. Being a **chartered accountant** and having experience in **internal audit**, the candidate is well-suited for roles in financial analysis, **internal control**, or accounting within the **finance department**. The candidate **must be recruited** for such roles.

Candidate 9

Name: Alexander Smith

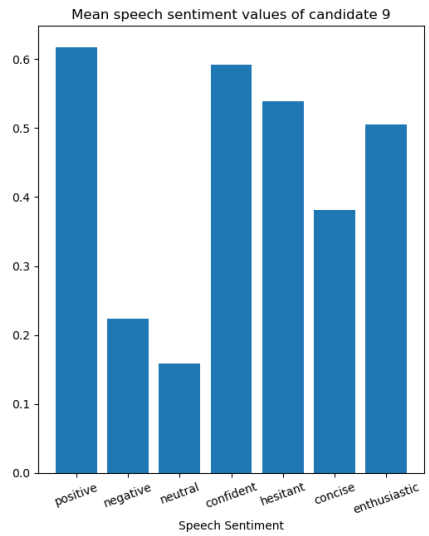
Education: B.Tech in Agriculture Engineering. M.Tech in Food Process Engineering.

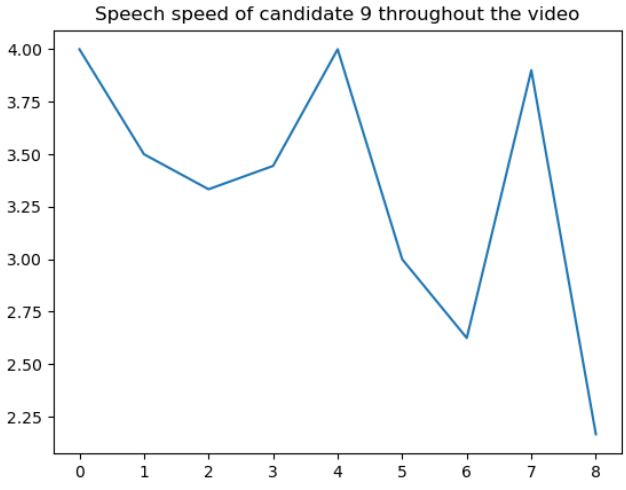
Work Experience: Co-founded an Agritech startup. Led a project on remote sensing IoT and AI in agriculture at an Agritech farm.

Relevant Skills: Entrepreneurship. Business development. Strategy.

Other details: Passionate about entrepreneurship and working with AI for societal development. Loves exploring challenging areas.

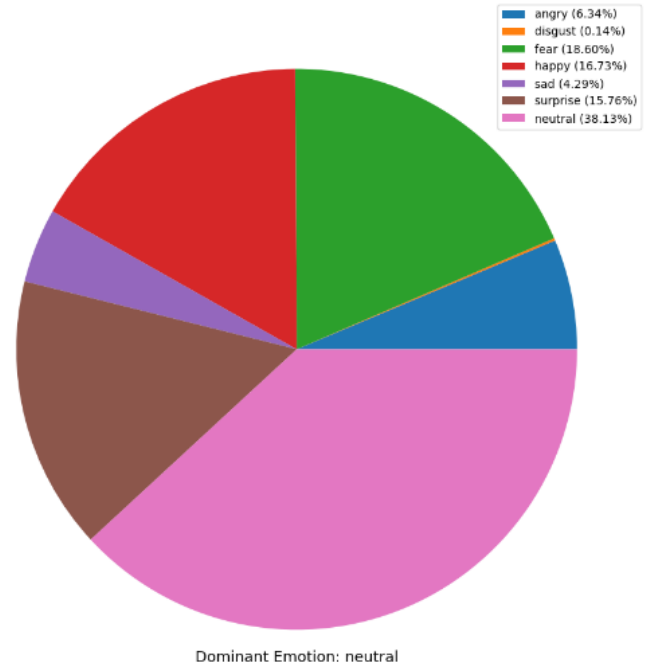
Language Proficiency: Good





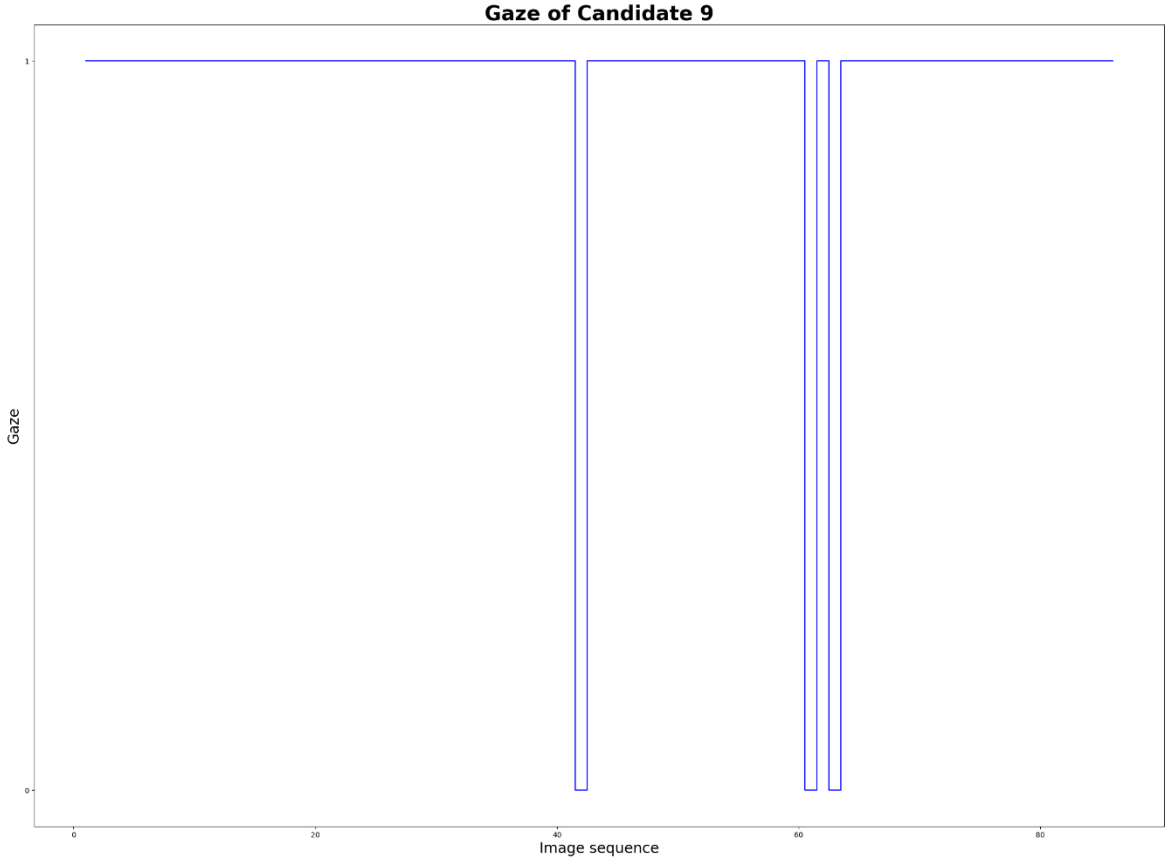
The plot on the left shows that the candidate is **positive, confident, and enthusiastic** through their words.

The plot on the right shows that a candidate is speaking at a **fluctuating pace**.



Emotions Distribution for Candidate 9

Gaze of Candidate 9

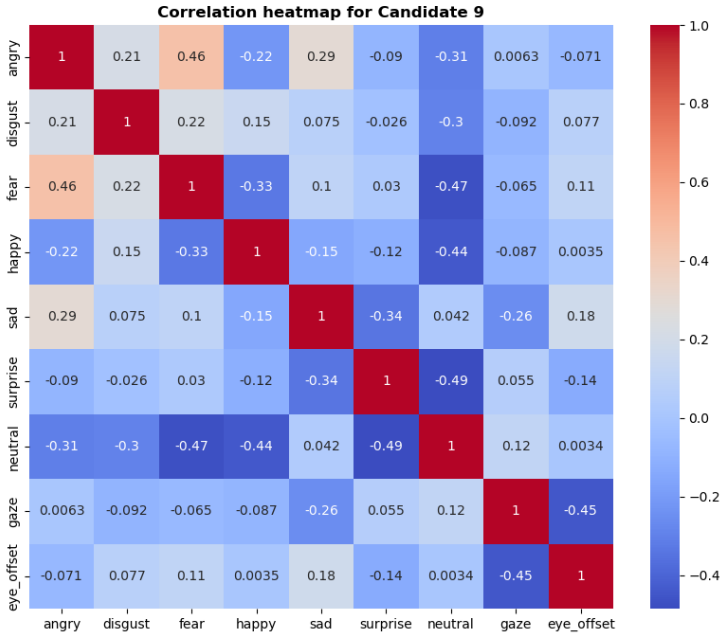


Percentage of frames looked into the camera: 96.51%

Gaze 1 = Person looked into the camera
Gaze 0 = Person did not look into the camera

The pie chart shows that the candidate is **fairly neutral** in his display of emotions.

The gaze step plot shows that the candidate **looks into the camera** for almost the entirety of the video.



The correlation heatmap does not provide any information as there are not any highly correlated values.

The candidate has a strong educational and technical background in food technology and agritech, showing a niche expertise. The candidate has an entrepreneurial mindset with experience in co-founding an agritech startup and leading technology-based agricultural projects. Apart from this he is positive, confident and enthusiastic in his words and emotions. With these skills and characteristics, the candidate would be a valuable resource in technology, research and development, and business development departments in any company and **must be recruited**.

Candidate 10

Name: Michael Ramos

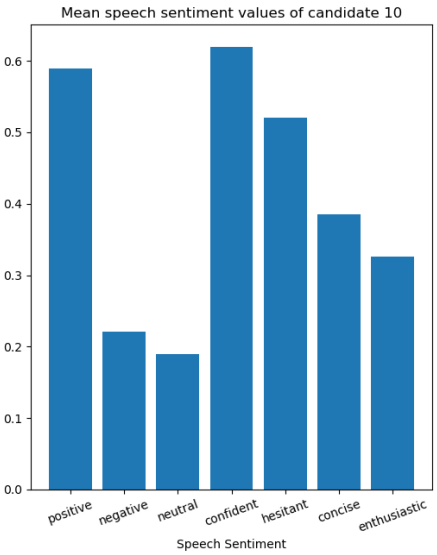
Education: B.Com Honours.

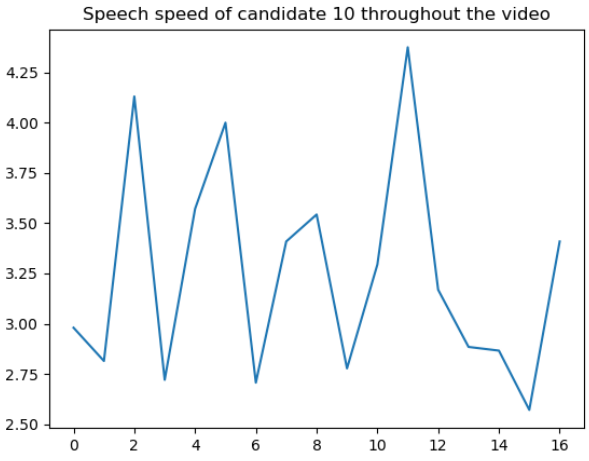
Work Experience: Internships as Accounting Associate and Tax Associate.

Relevant Skills: Accounting. Tax analysis. Leadership.

Other details: Leadership experience in student committees and social activities. Passionate about applying knowledge in real-life scenarios and creating value in the long term.

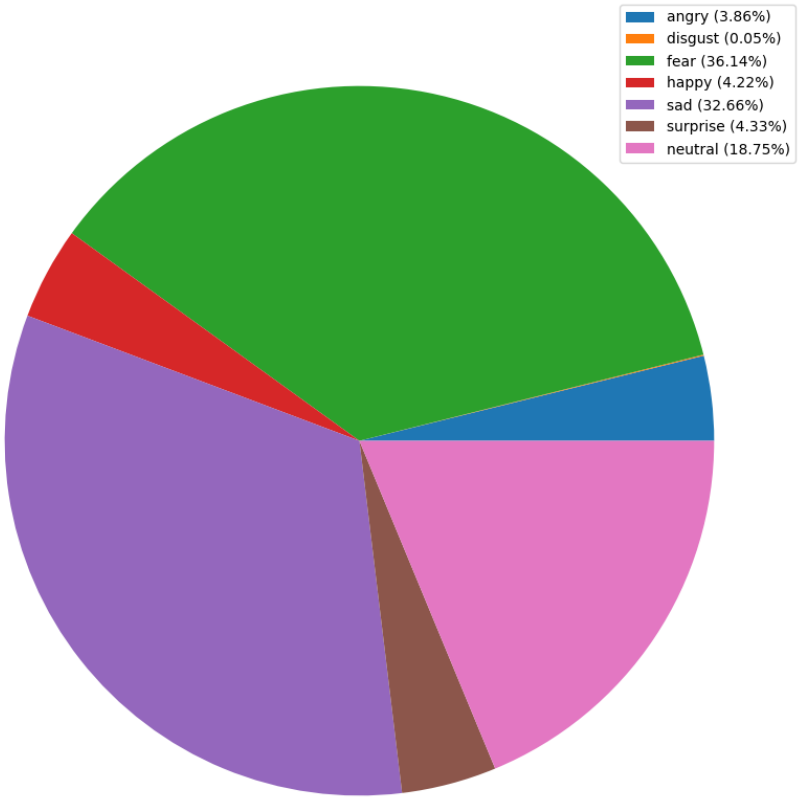
Language Proficiency: Good





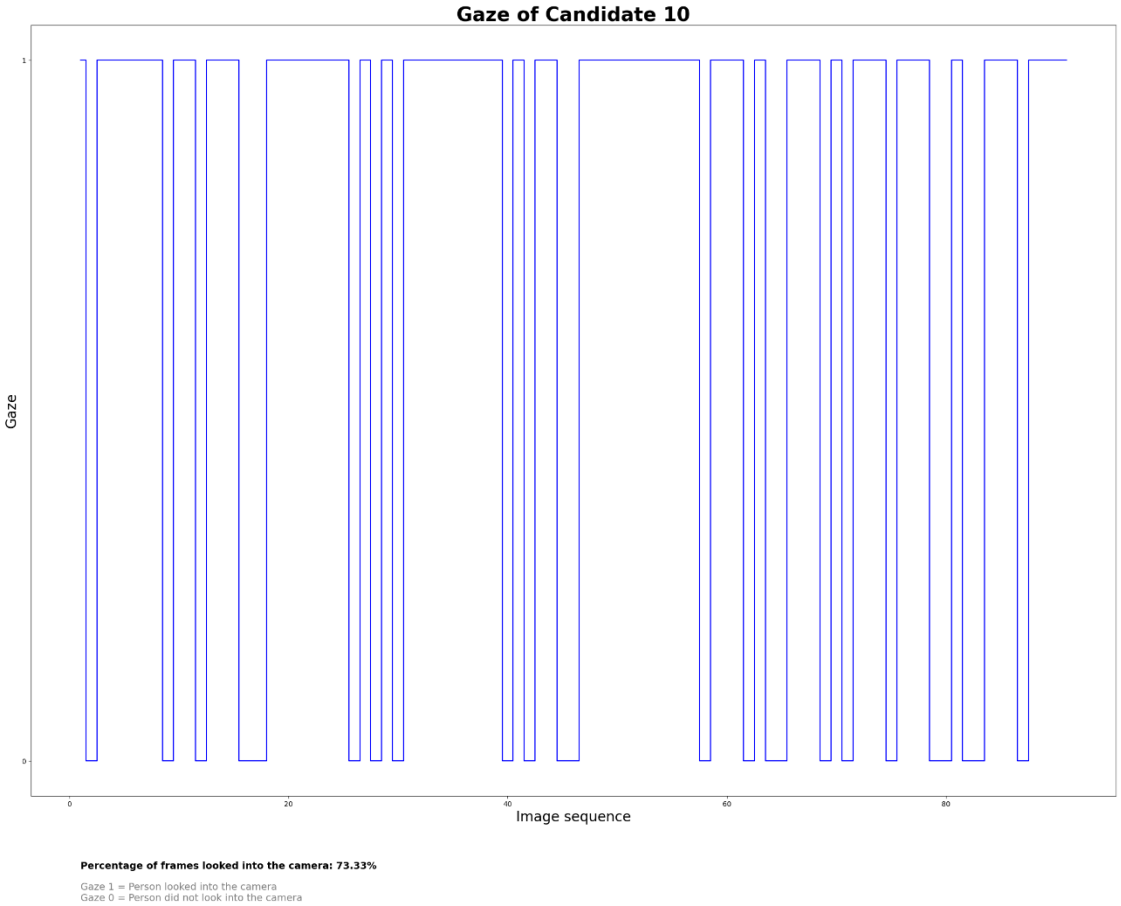
The plot on the left shows that the candidate is **positive and confident** through their words.

The plot on the right shows that a candidate is speaking at a **stable pace** of 3 – 3.5 words per minute.



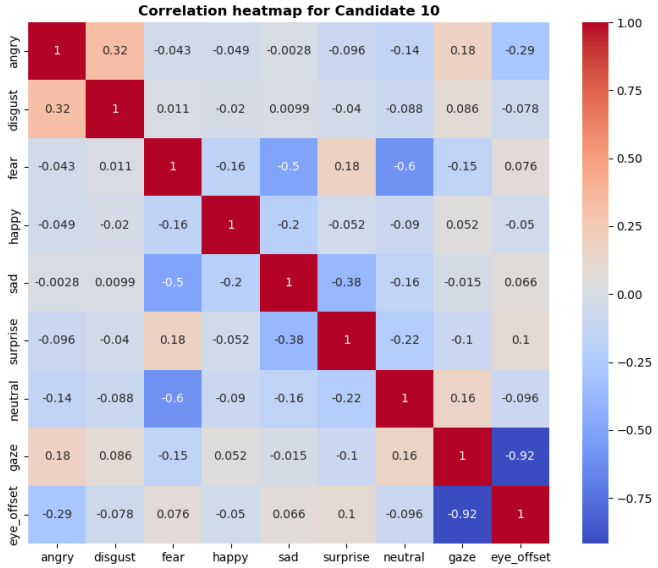
Dominant Emotion: fear

Emotions Distribution for Candidate 10



The pie chart indicates that the candidate majorly displays emotions of **fear or sadness**.

The gaze step plot shows that the candidate **looks into the camera** for majority of the video.



The heatmap shows **negative correlation between fear and neutral**, indicating that the candidate either shows emotions of fear or neutral emotions.

Although the candidate displays emotions of fear or sadness in the video, he exhibits positivity and confidence in his words. The candidate has educational and work experience that could prove useful in **accounting departments**. The candidate **could be recruited** for such roles.

Conclusion

In this study, we performed exploratory data analysis to get valuable information from the data. Our primary objective was to gain insights from the data provided that could be useful in deciding on the recruitment of the candidate.

The process consisted of the understanding the data, analyzing and visualizing the data and summarize the findings from the data analysis and the prompt engineering for each candidate.

For the purpose of data analysis and visualization, we used correlation heatmaps and different kinds of plots for different types of data.

Finally, we summarized the findings for each candidate and decided whether the candidate is suitable for recruitment or not.

It was finally found out that candidates 1, 6, 8, and 9 are candidates with high potential in their areas of expertise and must be given higher priorities during recruitment. Candidates 2, 4, 7 and 10 can be recruited for entry level or intermediate roles in their areas of expertise. These candidates do not have the best potential to be given higher priorities but they are enthusiastic and can turn out be good employees. Candidates 3 and 5 are not suitable for recruitment.