Advanced EDA of Video Text Dataset

Data Science Team

September 17, 2024

Dashboard

For seeing all the code live interactively,

Visit Dashboard

https://eda-analysis-iby-0.streamlit.app/

Contents

1 Introduction

This document presents an extensive Exploratory Data Analysis (EDA) of a video text dataset, focusing on emotional and communication attributes of students.

2 Data Overview

(a) Emotion Data Overview

emotion_data: This dataset contains the following columns. We can inspect the first few rows using *emotion_df.head()*.

Emotion Data Sample

This is a sample of emotion data. **Description:** This dataset represents the emotions detected at each timestamp of the video, along with the dominant emotion for each image sequence.

Image Seq	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral	Dominant Emotion
0	4.32	0.00	2.88	1.65	2.78	0.60	87.77	Neutral
1	53.23	2.98	12.74	1.52	1.05	27.22	1.26	Angry
2	8.80	0.03	2.97	16.83	39.88	0.28	31.21	Sad
3	9.45	0.11	1.55	20.93	3.50	0.91	63.54	Neutral
4	56.00	0.00	0.16	5.58	0.20	12.81	25.25	Angry

(b) Structure of Gaze Data

Gaze Data Structure

Image Seq	Gaze	Blink	Eye Offset
1	1	0	6.23
2	1	0	22.73
3	1	0	2.57
4	1	0	21.11
5	1	0	1.85

(c) Transcript Data Sample

Transcript Data Sample

id	text	tokens	positive	negative	neutral	confident	hesitant	concise	enthusiastic	speech_speed
0	Hello, I am J	[50364, 24]	0.5803	0.1523	0.2675	0.8467	0.8457	0.6358	0.6478	2.518
1	IIM Coikode.	[50642, 28]	0.5503	0.1893	0.2604	0.6793	0.7337	0.5441	0.4174	3.2178
2	Technology	[50844, 15]	0.6399	0.1111	0.2490	0.9027	0.8346	0.7159	0.7001	2.8689
3	of three yea	[51088, 29]	0.4419	0.3992	0.1589	0.7743	0.8130	0.5225	0.2799	3.750
4	as a medical	[51288, 38]	0.2363	0.5320	0.2317	0.2860	0.5614	0.3344	0.1973	3.5417

(d) Basic Statistics

Basic Statistics

	id	seek	start	end	positive	negative	neutral	confident	hesitant	concise	enthusiastic	speech_speed
count	18	18	18	18	18	18	18	18	18	18	18	18
mean	8.5	3,009.33	41.0022	45.9311	0.7092	0.1412	0.1496	0.7338	0.4852	0.4294	0.4665	3.1138
std	5.3385	2,598.47	26.117	26.2949	0.2073	0.1549	0.0810	0.2083	0.2608	0.2726	0.2863	0.600
min	0	0	0	5.56	0.2363	0.0050	0.0146	0.2860	0.0084	0.0128	0.0886	2.0349
25%	4.25	0	19.68	24.4	0.5879	0.0433	0.0829	0.5769	0.3429	0.2808	0.2114	2.6057
50%	8.5	2,776	40.56	46.64	0.7397	0.0804	0.1557	0.7899	0.4078	0.4415	0.4189	3.1342
75%	12.75	5,336	62.42	66.66	0.8701	0.1602	0.2246	0.8986	0.7108	0.6129	0.6870	3.5897
max	17	8,272	82.72	88.72	0.9804	0.5320	0.2675	0.9809	0.8457	0.9197	0.9903	4.1667

Note: In the transcript data, the columns *text*, *tokens*, *temperature*, *avg_logprob*, *compression_ratio*, and *no_speech_prob* are removed for simplification.

3 Data Preprocessing

Steps for Data preprocessing

The following steps were performed to preprocess the data before conducting any Exploratory Data Analysis (EDA):

- **Handling Missing Values:** Replaced missing or null values using techniques such as mean, median, or mode imputation based on the feature type.
- **Encoding Categorical Features:** Converted categorical variables into numeric form using label encoding or one-hot encoding.
- **Scaling and Normalization:** Applied standard scaling or min-max normalization to ensure that features with different units or ranges do not dominate the model's learning process.

4 Data Preparation and Integration

So, the data is spread across three different CSV files, and what I need to do is pull out key features from all of them to create a single, clean DataFrame.

First off, I'm going to take the **dominant emotions** for each student across the entire video. To do this, I'll average the dominant emotions over every frame and then store them as **dominant_emotion_top_1** and **dominant_emotion_top_2**. These two values will give us a quick idea of the main emotional tones for each student.

1. First, I'll **extract and compute the average values** from the **transcript data**. This is important because we need to analyze how students behave by looking at things like **positivity**, **negativity**, and **speech speed**. By averaging these features across their complete transcript, I can see overall patterns, like how often they seem confident or hesitant.

2. Then, I'll **store these averages in a list of DataFrames**. The reason for this is to keep track of each student's data separately at first. This way, when I combine them later, it's easier to manage and analyze.

3. After that, I'll **combine all the averages into one final DataFrame**. This step is key because it gives me a **well-organized dataset** where I can quickly compare each student's average speech features. So, up to here, the **transcript part** is done.

4. Next, I'll **extract the dominant emotions** from the **emotion data**. The goal here is to capture the emotional tone of each student by identifying the **most frequent emotions**—whether they're happy, sad, or something else. This adds more depth to the average speech features we just looked at.

5. Once I have the dominant emotions, I'll **store the top two** for each student. This is important because it gives us a quick snapshot of their **emotional profile** during speech, covering both the primary and secondary emotions. It's like getting the full emotional picture.

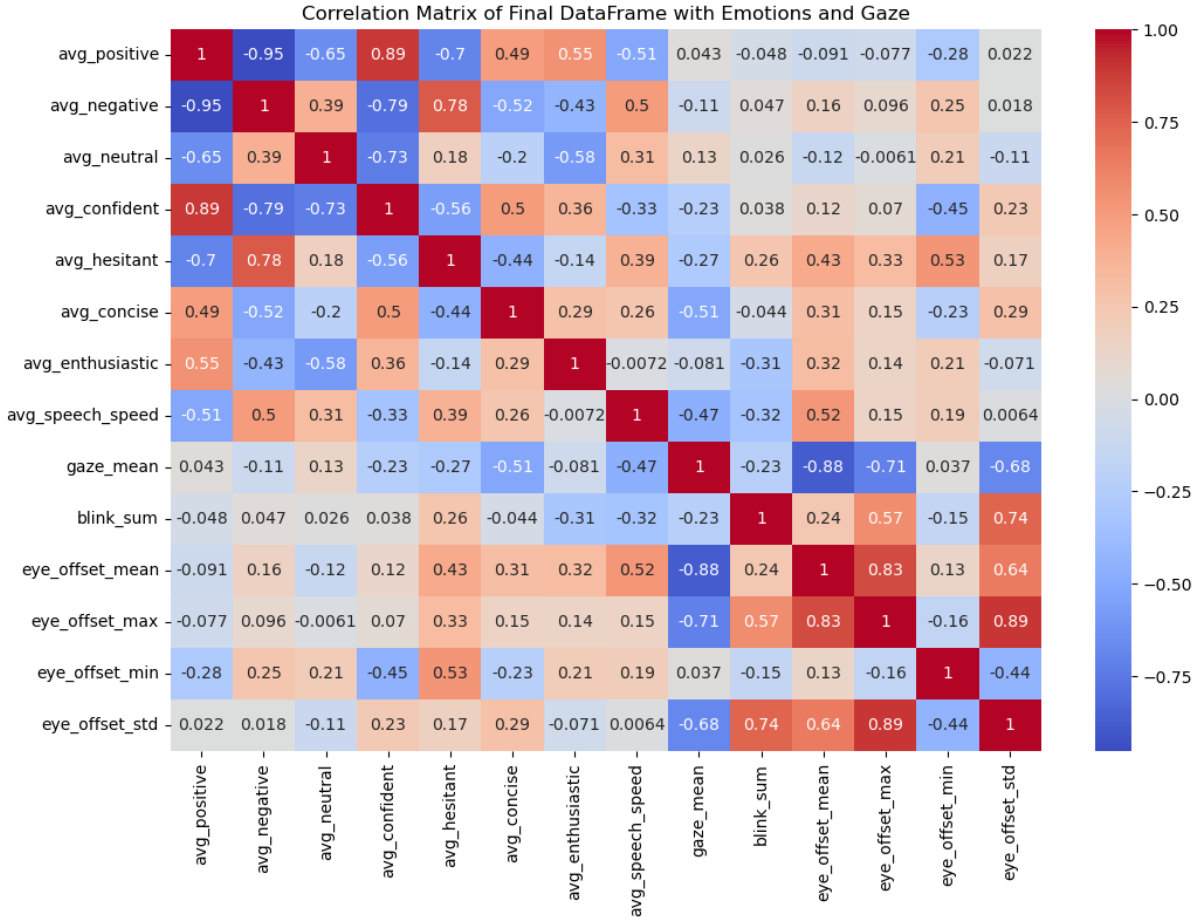
6. I'll then **create a DataFrame** to hold all these dominant emotions. Having this in a separate DataFrame makes it easy to **merge with the average speech features**, so I can analyze everything together in one place.

7. Now, I'll **merge the average features DataFrame with the dominant emotions DataFrame**. This combination gives me a comprehensive dataset, mixing both the **quantitative data** (like positivity and speech speed) with more **qualitative data** (like emotions), which lets me dig deeper into how students' behavior and emotions interact.

8. Finally, I'll **display the complete DataFrame**. This final view will give a **clear overview** of each student's **speech patterns** and **emotional tendencies**, which will help understand their **communication behavior** better and could guide future feedback or interventions.

So this is my new DATAFRAME , Let's call it FINAL_df

5 Advanced Analysis



Insights from the correlation matrix

Focusing on this image, We can conclude :

- **avg_positve and avg_confident is highly correlated as red means they are correlated.**
- **avg_hesitant and avg_negative is also correlated as they are blue and their correlation score is 0.77(quite high) .**
- **avg_enthusiastic and avg_confident is also correlated though not as high as avg_positive and avg_confident.**
- **avg_negative and avg_confident is highly uncorrelated and their correlation score is -0.73(quite high) .**

In this way, we can see the correlation between the features like which features are dependent or which are not. From this we can if a student whose text content score is positive, then he/she is more likely more confident and enthusiastic in comparison to the student whose text content score is negative. This fact will help in further analysis.

Distribution plots were generated for various features to understand their spread and central tendencies. For example:

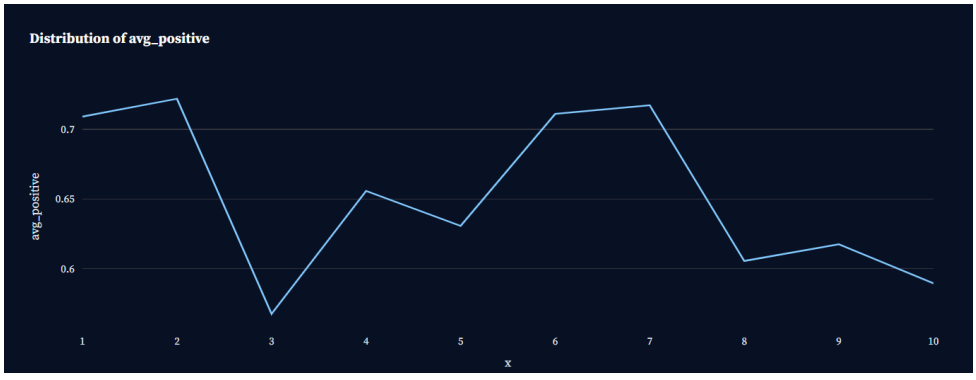


Figure 1: Distribution of avg_positive scores

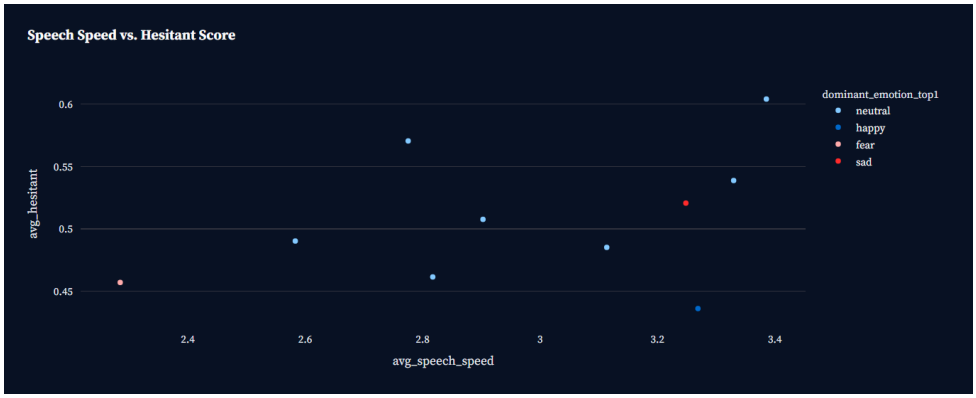
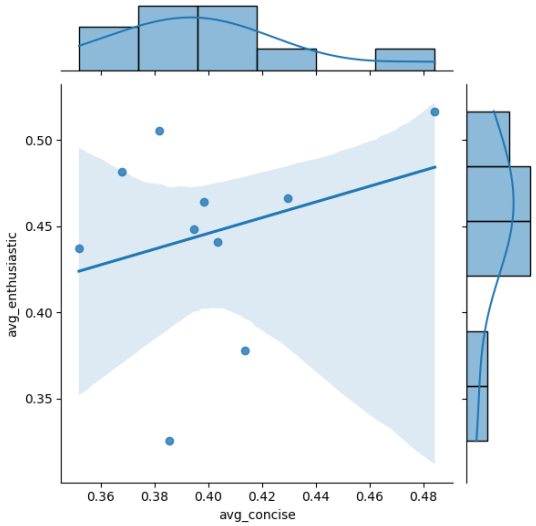


Figure 2: Speech Speed vs. Hesitant Scores

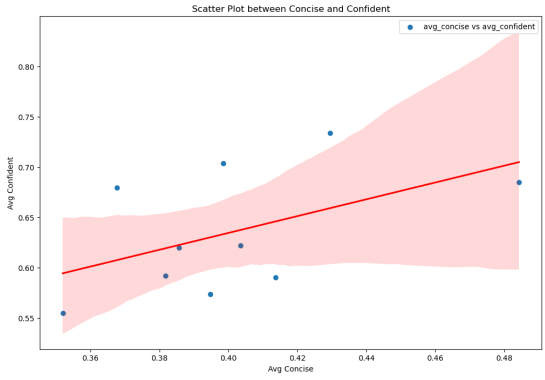
(a) Communication Skills Analysis

- **First**, I will investigate the relationship between **conciseness** and **enthusiasm** of the students.
  For this, I am plotting a joint plot between avg_concise and avg_enthusiastic. From the plot, we can see that there is a positive correlation between the two features.

  This indicates that students who are more concise in their speech are also more enthusiastic.



Joint plot between avg_enthusiastic and avg_concise.

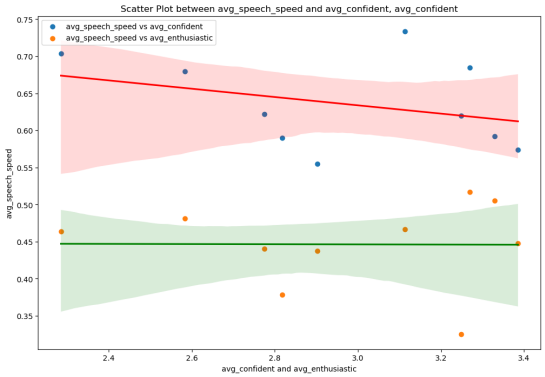


Scatter plot between avg_speech_speed and avg_hesitant.

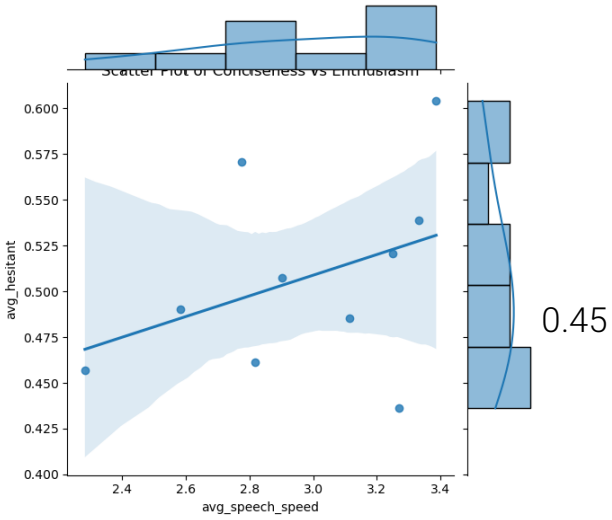
Comparison of conciseness vs enthusiasm and confidence.

- **Communication** skill is also dependent on the **speed of speech**. To analyze this, I am plotting a scatter plot between avg_speech_speed and avg_hesitant.

The plot reveals a negative correlation between these two features, meaning that students who speak faster are less hesitant in their speech.



Scatter plot between avg_speech_speed and avg_confidence.

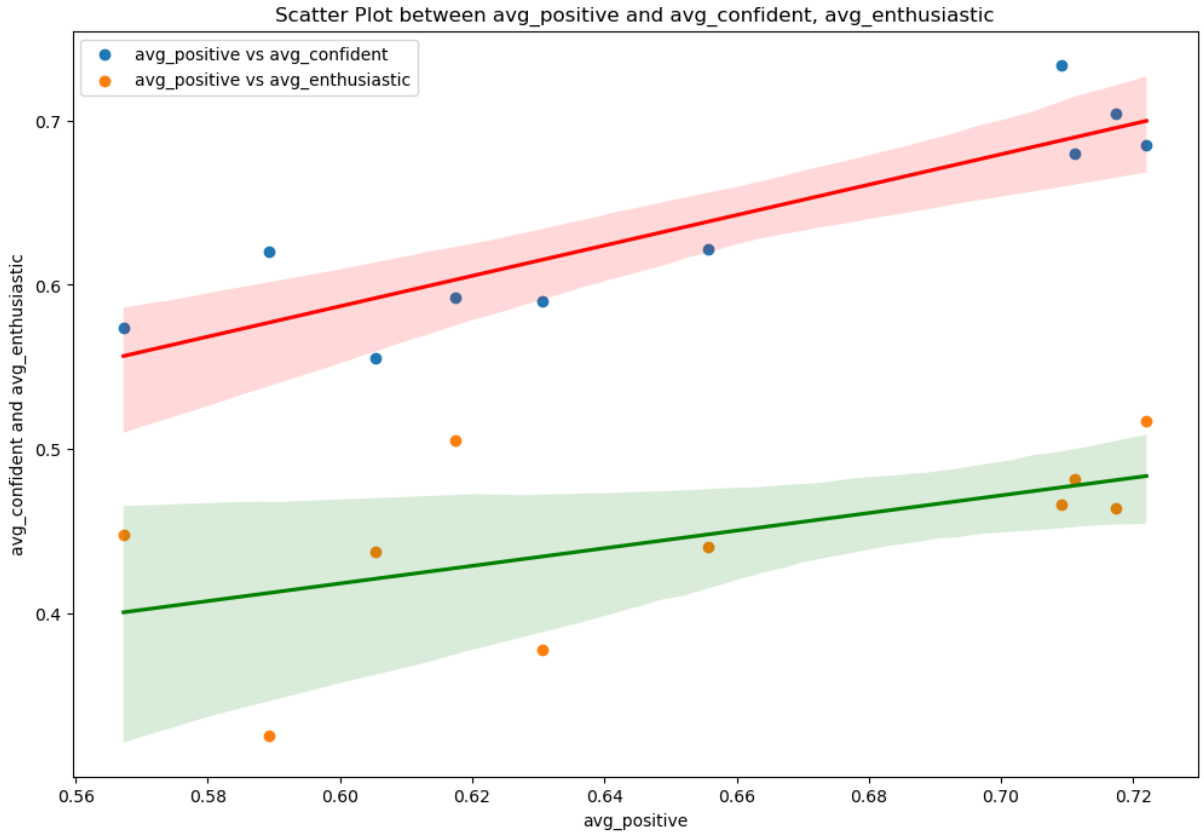


0.45

Scatter plot between avg_speech_speed and avg_hesitant.

- **Text** content scores (positive, negative, neutral) are also crucial in communication skills. Therefore, I will analyze the relationship between **positivity** and **confidence** of the students.

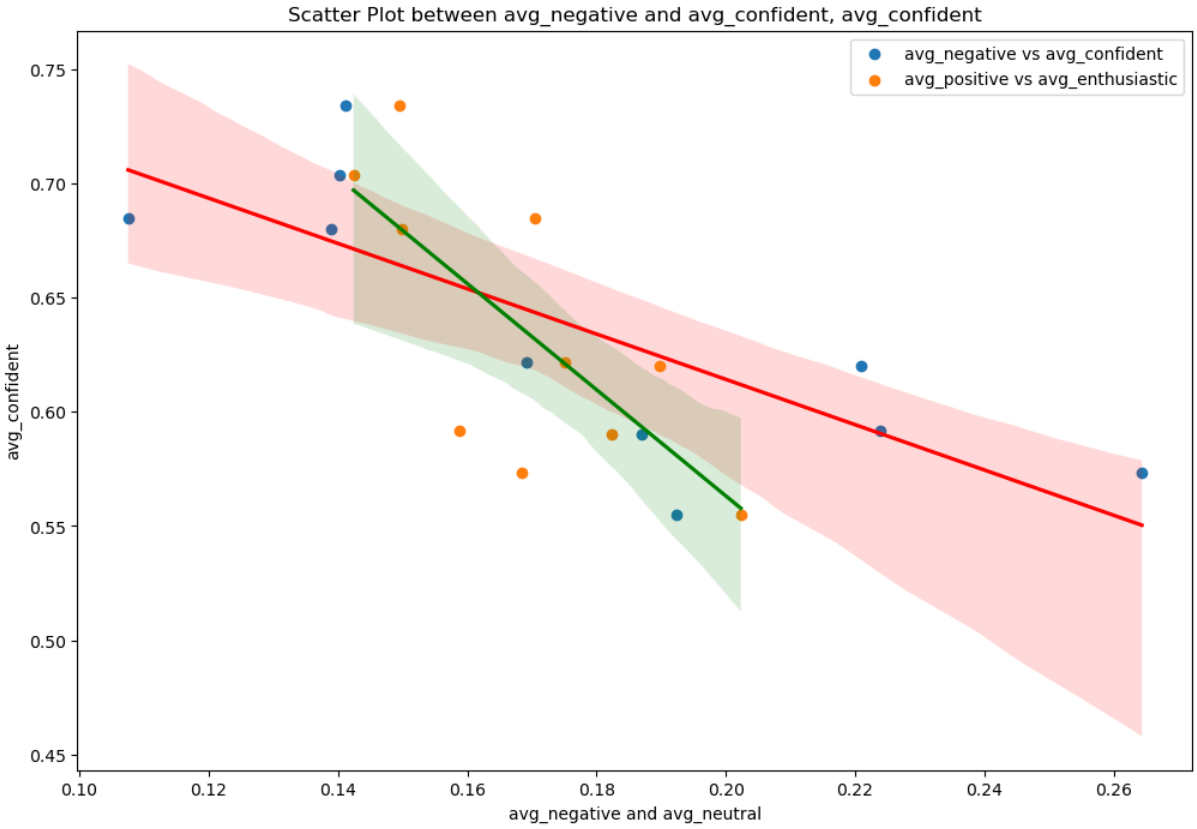
  For this, I am plotting a scatter plot between avg_positive and avg_confident. The plot shows a positive correlation between these features, indicating that students with a more positive text content score are also more confident in their speech.

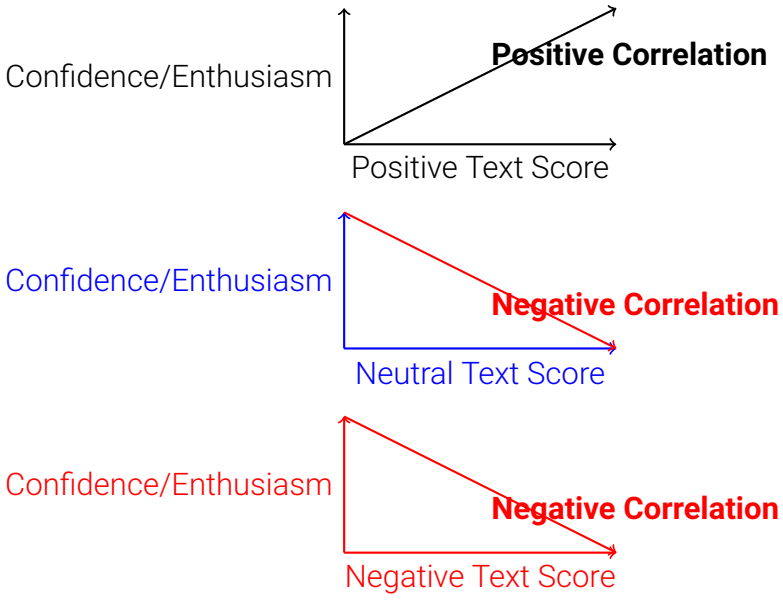


From this graph, we can clearly see the linear relation between avg_positive vs avg_confident and avg_positive vs avg_enthusiastic.

- **Additionally**, the relationship between avg_negative and avg_confident is also linear but in the opposite direction.

  This implies that students with a more negative text content score tend to be less confident and enthusiastic.





Correlation Diagram
for Text Scores with Confidence and Enthusiasm

- An interesting insight from the above graph is that avg_neutral is also linearly related to avg_confident and avg_enthusiastic, but in the opposite direction. This provides new insights into how neutrality in text content affects communication skills.

-

(b) Emotional State and Body Language Analysis

1.**Emotional Stability Analysis**
So, I have calculated the **emotional stability** of each student by analyzing the **variability** in their emotions throughout the video.

Emotional stability is a key factor in understanding how well students can manage their emotions during communication.
This is important because it helps us understand how consistent a student is in expressing different emotions.

Warning

Since there are 10 students and in this report, for a sample, I am showing the analysis for one student only. For other students' analyses, kindly visit the full report here.

There is slight moment in between where the student is showing fear and sadness.

This indicates that he is able to maintain a consistent emotional tone (which is neutral) during communication, which is a positive trait for effective interaction.
`This student has shown a high level of emotional stability throughout the video, with minimal fluctuations in their emotional state.`

`There is slight moment in between where the student is showing fear and sadness.`

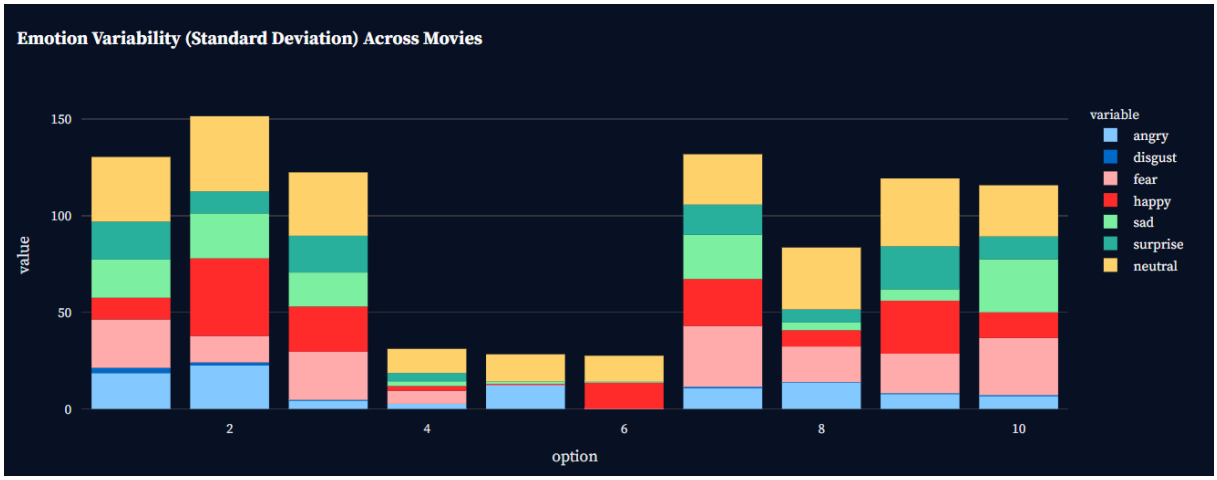
`This indicates that he is able to maintain a consistent emotional tone (which is neutral) during communication, which is a positive trait for effective interaction.`

2. **Emotion Variability Analysis**

Emotion variability is another important aspect of emotional intelligence, as it reflects how well students can adapt their emotions to different situations.

The amount of emotions a student is showing during their speech can be an indicator of their emotional intelligence.

If he shows fear or sadness for a long time, then it can be a sign of less communication skills and poor body language.



Like in this graph, we can see that except student 5 and 6, all students are showing a good amount of variability in their emotions.

3. **Body Language Analysis**

So for this, I made a new dataframe named fiNAL-GAZE_DF which contains the gaze data of all the students.

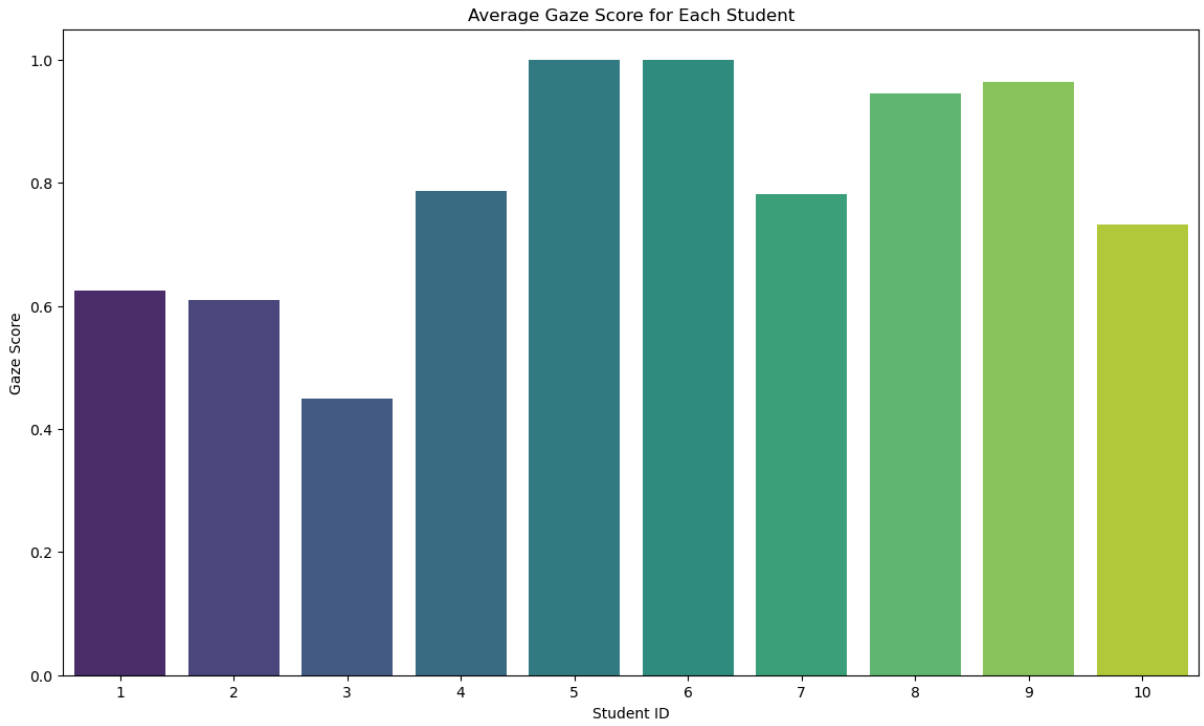
It contains the following columns:

- `movie_id:` movie_id
- `Gaze_score:` The gaze score of the student: proportion of time the candidate spends looking at the camera..
- `blink_sum:` The blink sum of the student.
- `eye_offset_std:` The eye offset standard deviation of the student.

1. If the standard deviation (std) of the eye offset of a person in a video is too high, it suggests that the person's gaze is not stable or consistent across frames

i. **Gaze Analysis**

Gaze is an important aspect of body language that can reveal a lot about a student's focus and engagement during communication.



BASED ON THE ABOVE GRAPH, FOLLOWING OBSERVATIONS CAN BE MADE:

1. Highly Engaged (0.85 - 1.0): Students who maintained frequent or constant eye contact with the camera.
   In this category, Student 5.6,8,9 fall as their gaze score is above 0.85.
2. Moderately Engaged (0.7 - 0.85): Students who maintained moderate eye contact with the camera.
   In this category, Student 4,7,10 fall as their gaze score is between 0.5 to 0.85.
3. Low Engagement (Below 0.7): Students who maintained low eye contact with the camera.
   In this category, Student 1,2,3 fall as their gaze score is below 0.5.

ii. **Blink Analysis**

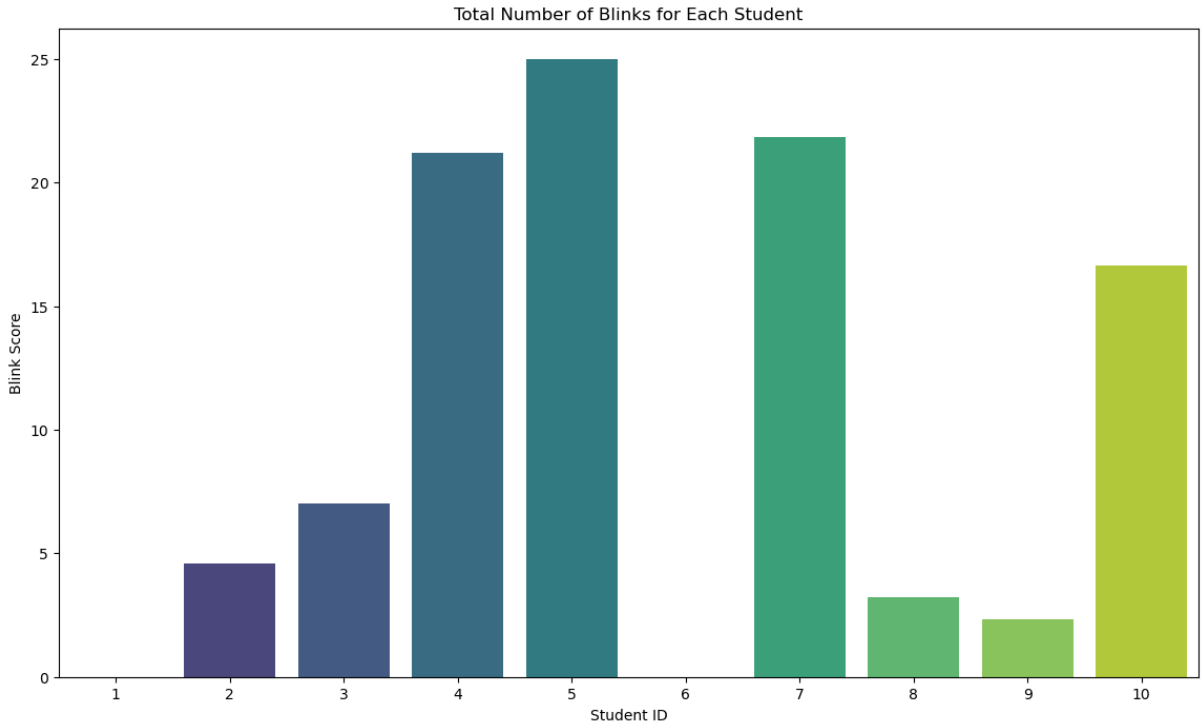
Blinking is another important aspect of body language that can indicate a student's level of comfort and confidence during communication.

If a student blinks too frequently, it may suggest nervousness or discomfort, while infrequent blinking may indicate confidence or focus.

Catch

Since the blinking rate will depend on the amount of time of video, i have to divided the blink$_sumwithtotal frames of the video to get the blink rate.$

Blink Rate = blink_sum / total_frames

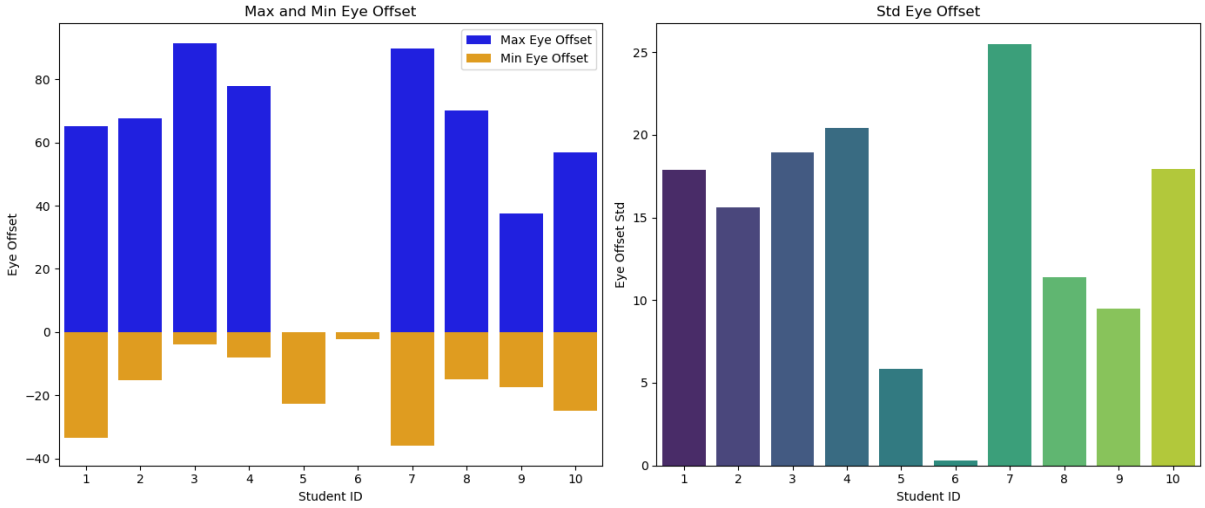


BASED ON THE ABOVE GRAPH, FOLLOWING OBSERVATIONS CAN BE MADE:

1. High Blink Rate: Students who blinked frequently during the video, i.e they are either nervour or feeling anxiety.
   In this category, Student 4,5,7,10 fall as their blink rate is .
2. Normal Blink Rate: Students who blinked at a moderate rate during the video.
   In this category, Student 1,2,3,5,6,8,9 fall as their blink rate is between 0.2 to 0.5.

iii. **Eye Offset Analysis**

Eye offset is a measure of how much a student's gaze deviates from the camera during communication.
So the higher the eye offset, the more the student's gaze is wandering away from the camera.



1. Max and Min Eye Offset (Left Plot):

This chart displays the maximum and minimum eye offsets for each student. The blue bars represent the max positive deviation, and the orange bars represent the max negative deviation.

2. Eye Offset Standard Deviation (Right Plot):

This chart shows the standard deviation of eye offsets for each student. The standard deviation indicates how much the student's eyes typically deviated from the mean eye position over the duration of the video.

BASED ON THE ABOVE GRAPH, FOLLOWING OBSERVATIONS CAN BE MADE:

1. `Highly Erratic Eye Movements:`
   - Student 7 shows the greatest overall deviation in both positive/negative offsets and in the standard deviation, indicating highly variable and erratic eye movements.
   - Students 3 and 4 also show significant deviation in eye movements, suggesting frequent or large fluctuations in where they were looking.
2. `Moderate Eye Movements:`
   - Students 1, 2, and 10 demonstrate moderate eye movement variability. They have noticeable deviations but are less erratic compared to students like 7 and 3.
3. `Stable Eye Movements:`
   - Students 5, 6, 8, and 9 exhibit the most stable eye movements, with minimal deviation from the mean eye position. This suggests that they maintained consistent eye contact with the camera throughout the video.

6 Individual Student Analysis

On the basis of above analysis, I have created a detailed analysis for each student.
His expertise, communication skills, emotional intelligence, and body language are analyzed in detail.

(a) Student 1

(b) Student 2

7 Statistical Analysis

8 Feature Importance

9 Conclusion

This comprehensive analysis of the video text dataset has revealed several significant patterns and relationships among emotional and communication attributes of students:

Strong correlation between confidence and positive emotions

Statistically significant difference between confident and hesitant scores

No significant relationship between positive emotions and speech speed

Hesitance and enthusiasm as the most important predictors of speech speed

These insights provide valuable information for understanding student behavior and can be utilized to optimize educational strategies, particularly in the context of video-based learning and assessment.

10 Future Work

To further enhance this analysis, consider the following directions for future research:

Conduct longitudinal studies to track changes in student communication patterns over time

Develop machine learning models to predict student performance based on communication attributes

Investigate the impact of different teaching methodologies on student communication styles

Expand the dataset to include a more diverse range of students and educational contexts

By pursuing these avenues, researchers and educators can gain even deeper insights into student behavior and develop more effective, personalized learning strategies.