Big Data Processing

Analysis of debates from the
2024 Portuguese election

Project done by group 1

Summary

1 - Exploratory Data Analysis (EDA)

2 - Data Cleaning

3 - Task 1 - Approach + Results

4 - Task 2 - Approach + Results

5 - Overall Results

Chapter 1 - Exploratory Data Analysis

Before working with the data, we need to understand its quality and what conclusions we can draw from it.

Objectives:

Understand Data Structure

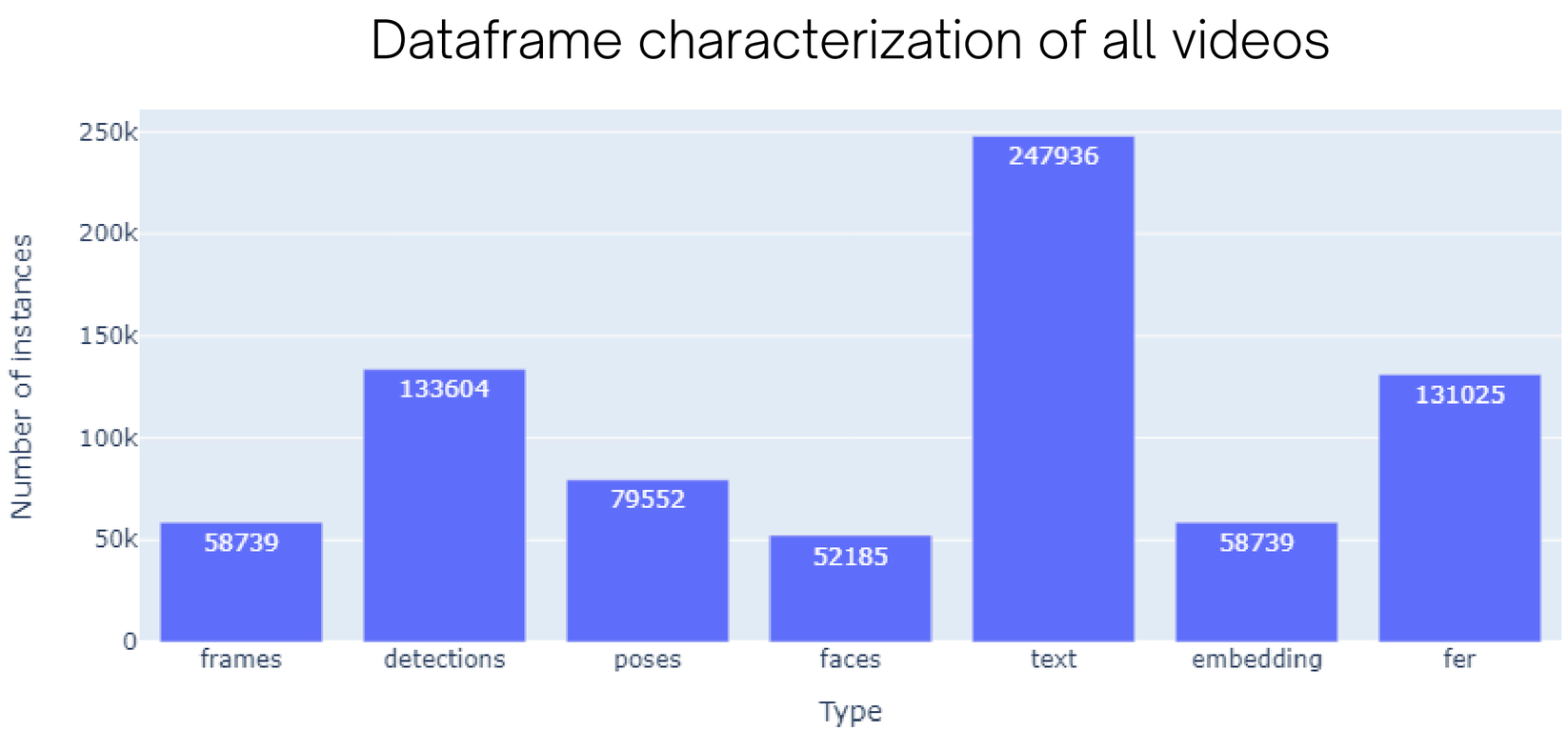
Assess Data Quality

Generate Initial insights

Data Structure

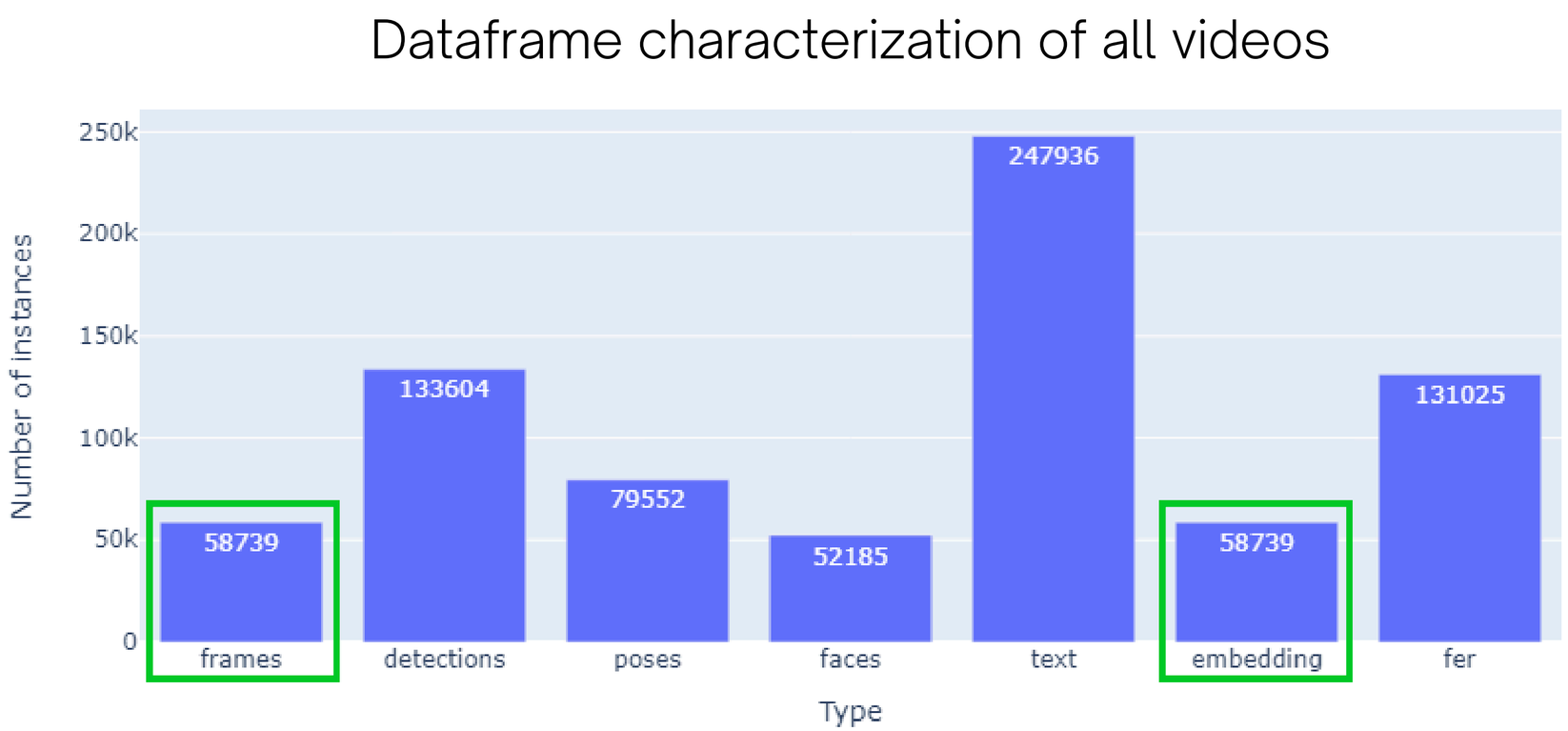
	filename	detections	poses	faces	text	embedding	fer
0	...	...	...	...	...	...	...
1	...	...	...	...	...	...	...
2	...	...	...	...	...	...	...

Data Structure



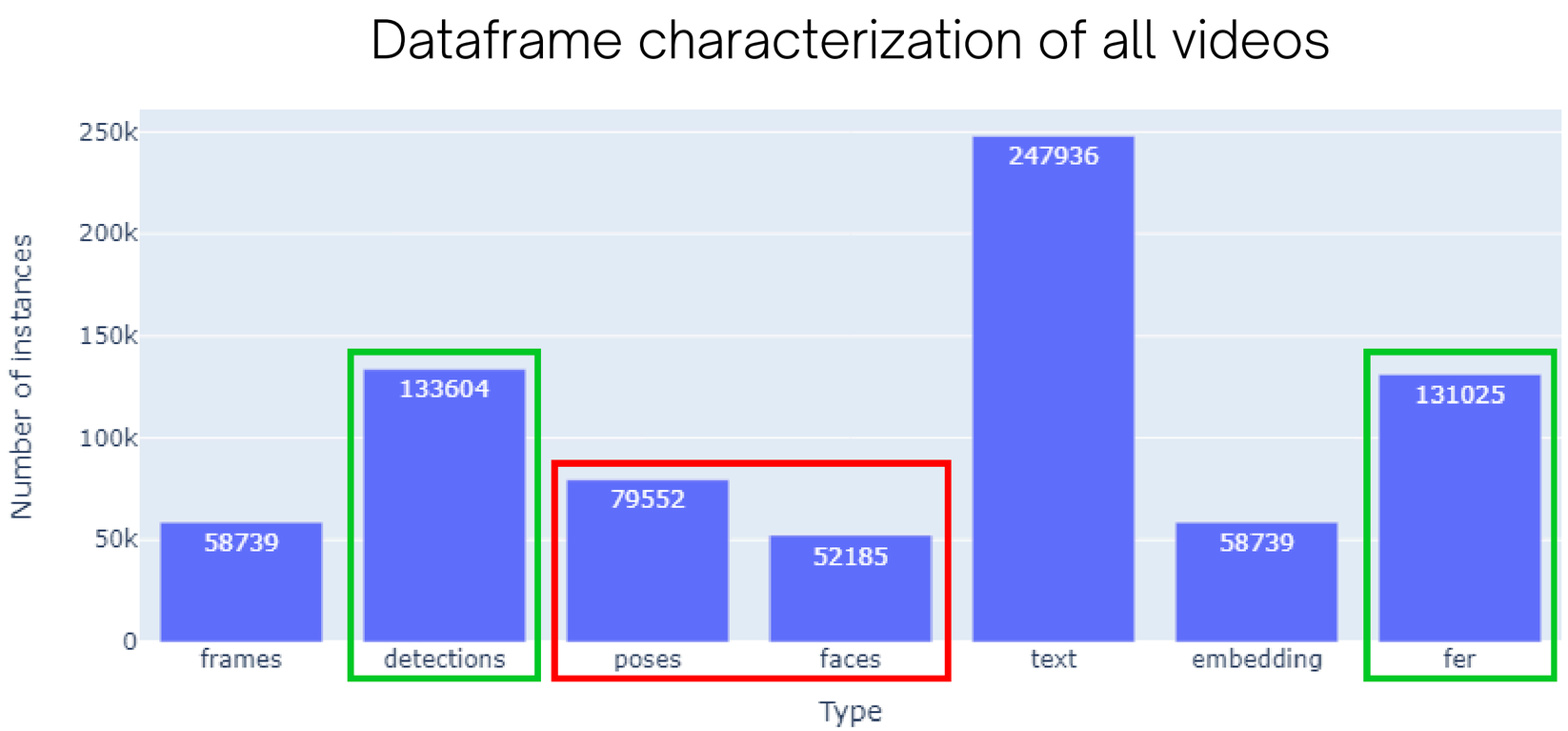
- The dataset provides a wealth of information to analyze.
- We need to select the most relevant data features to effectively accomplish the tasks.

Data Structure



- For each frame, there is an assigned image embedding vector.

Data Structure

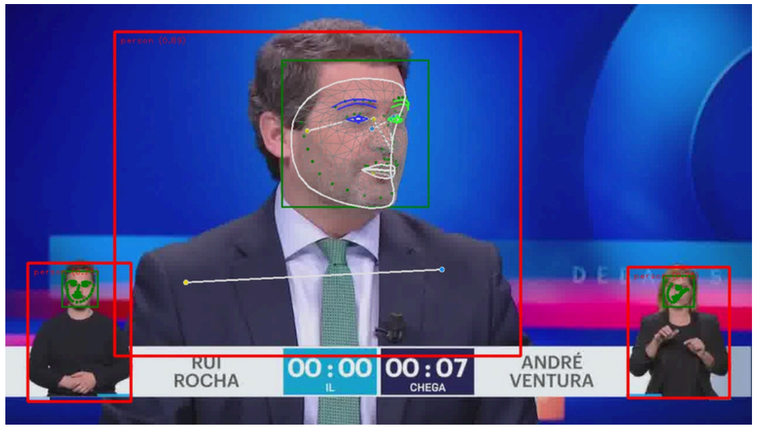


- 'detections' and 'fer' appear with roughly the same frequency.
- 'Poses' and 'faces' occur less consistently.

Data Quality

Frames

Frame not directly related to the debate



Frame with two interperters and screen overlay

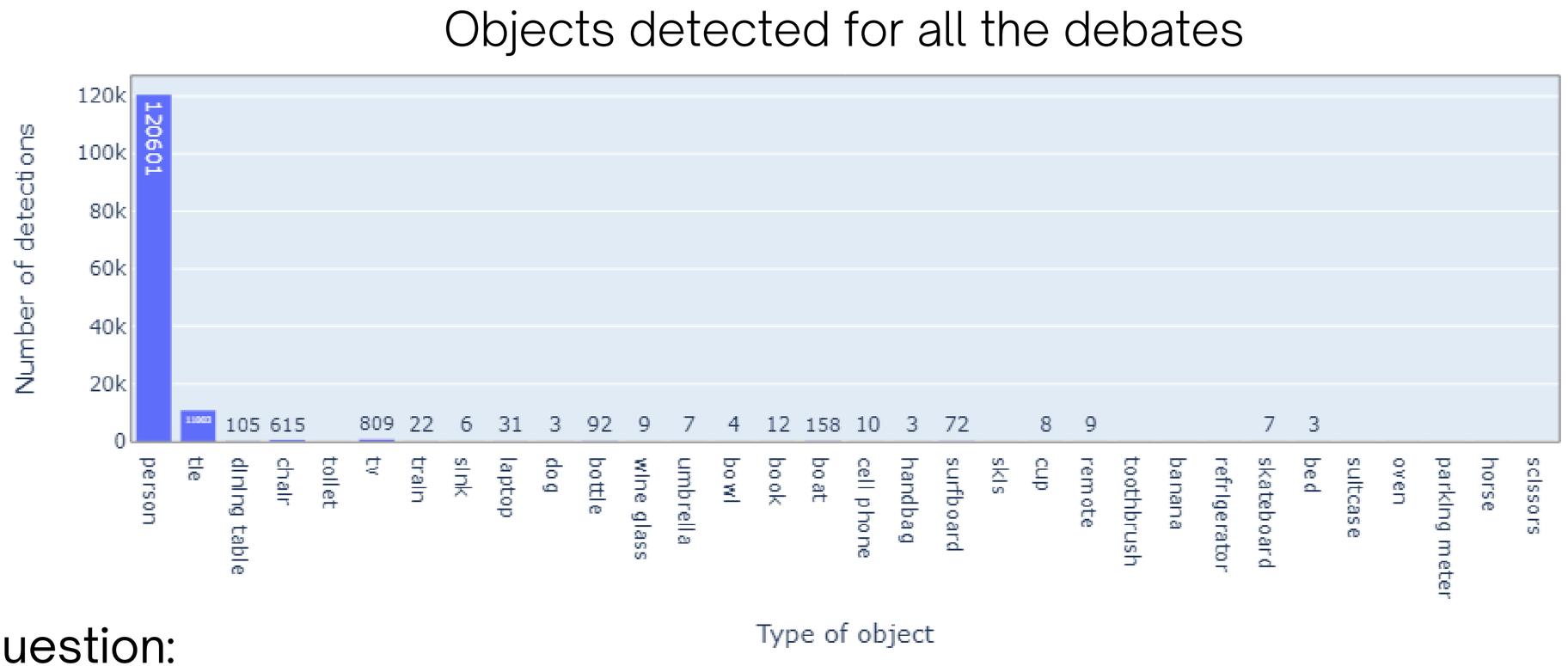
Frame without interperters or screen overlay

To ensure high-quality analysis, it is necessary to:

- Filter out frames that are not related to the debate.
- Distinguish interpreters from politicians and moderators.
- Address and manage inconsistent data.

Data Quality

‘detections’ objects

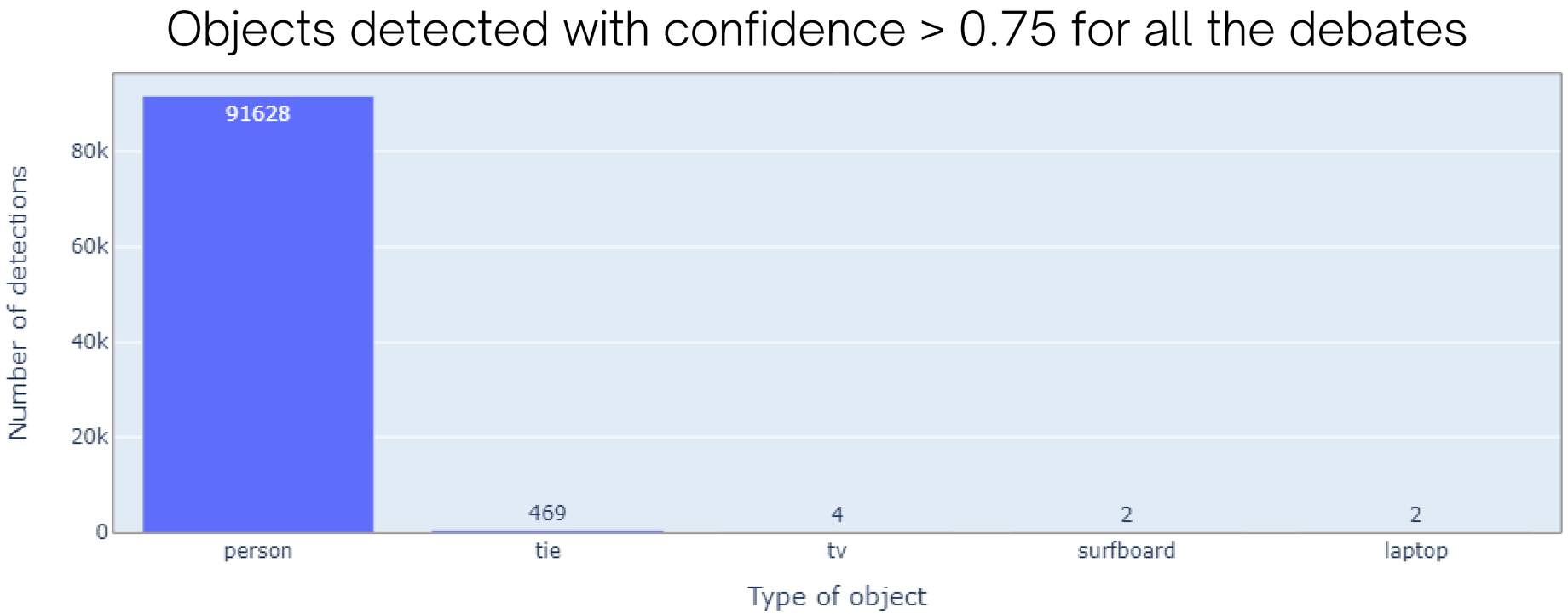


Question:

- Is the ‘object detection’ only useful to analyse persons?

Data Quality

‘detections’ objects



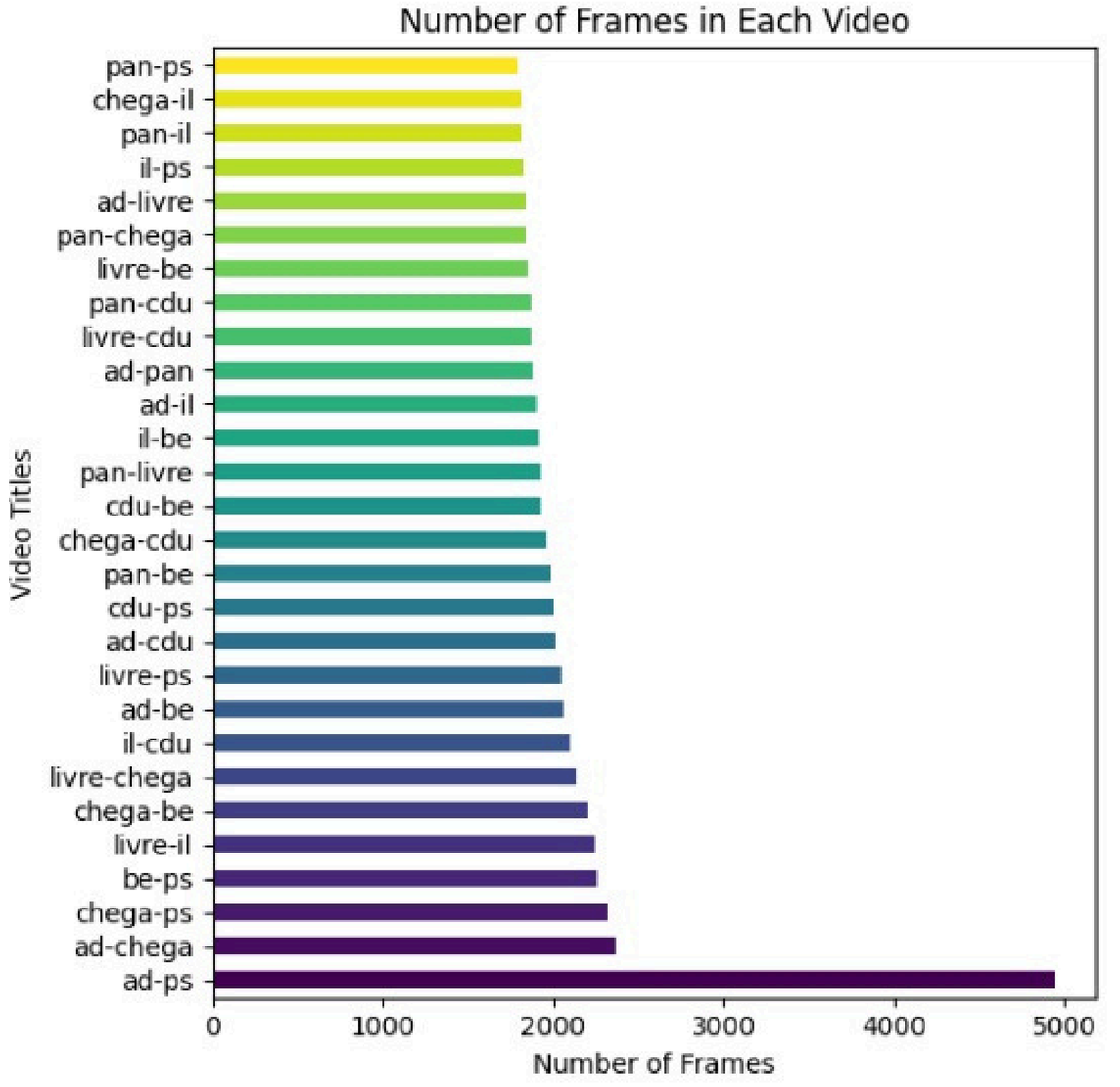
Conclusion

- 99,5% of the objects detected are the class ‘persons’.

Initial Insights

Video duration

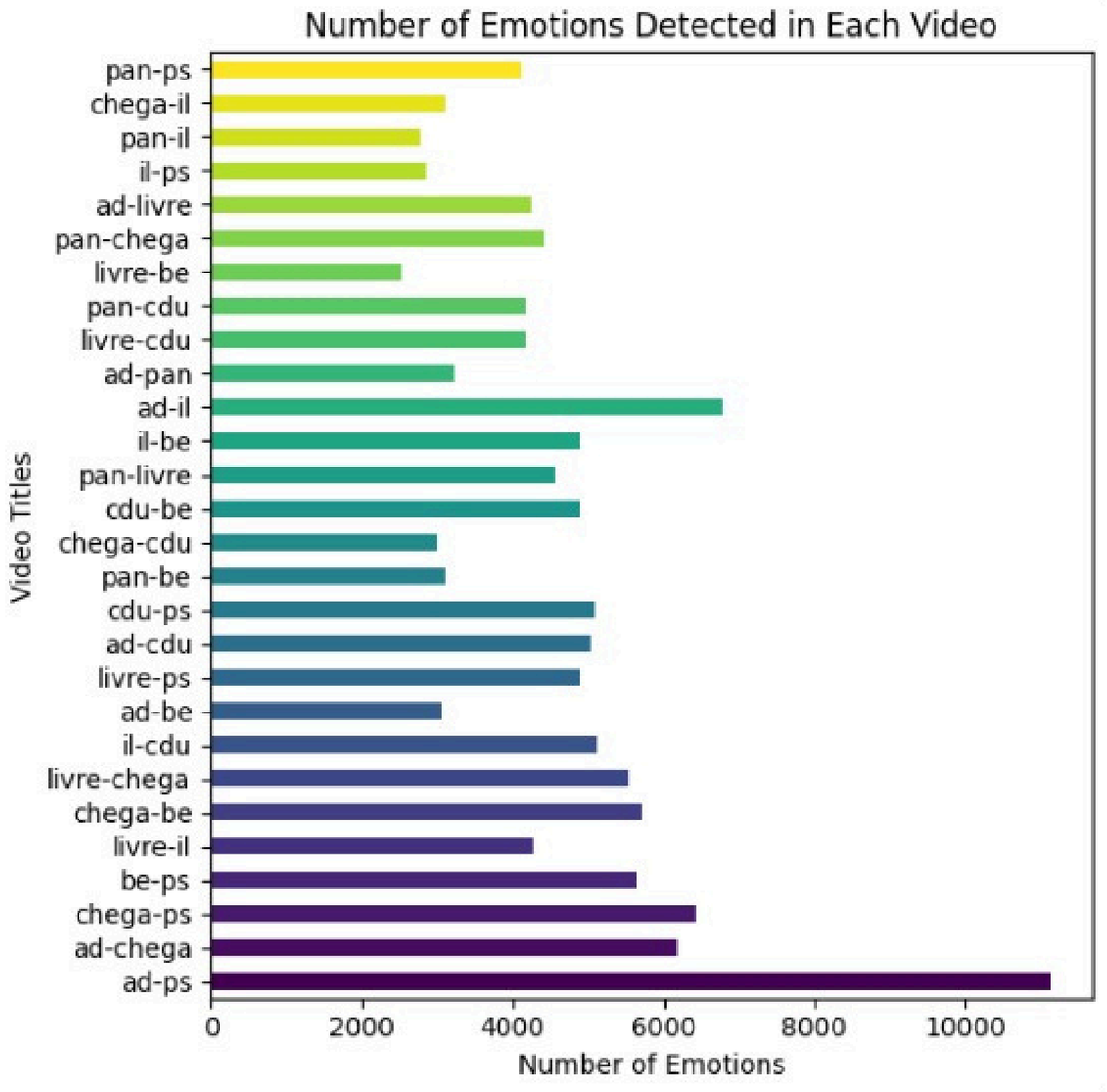
All the debates have roughly the same duration (~2,000 frames), except for ‘ad-ps,’ which has 5,000 frames.



Initial Insights

Emotions

The number of emotions detected is not directly correlated with the video duration.

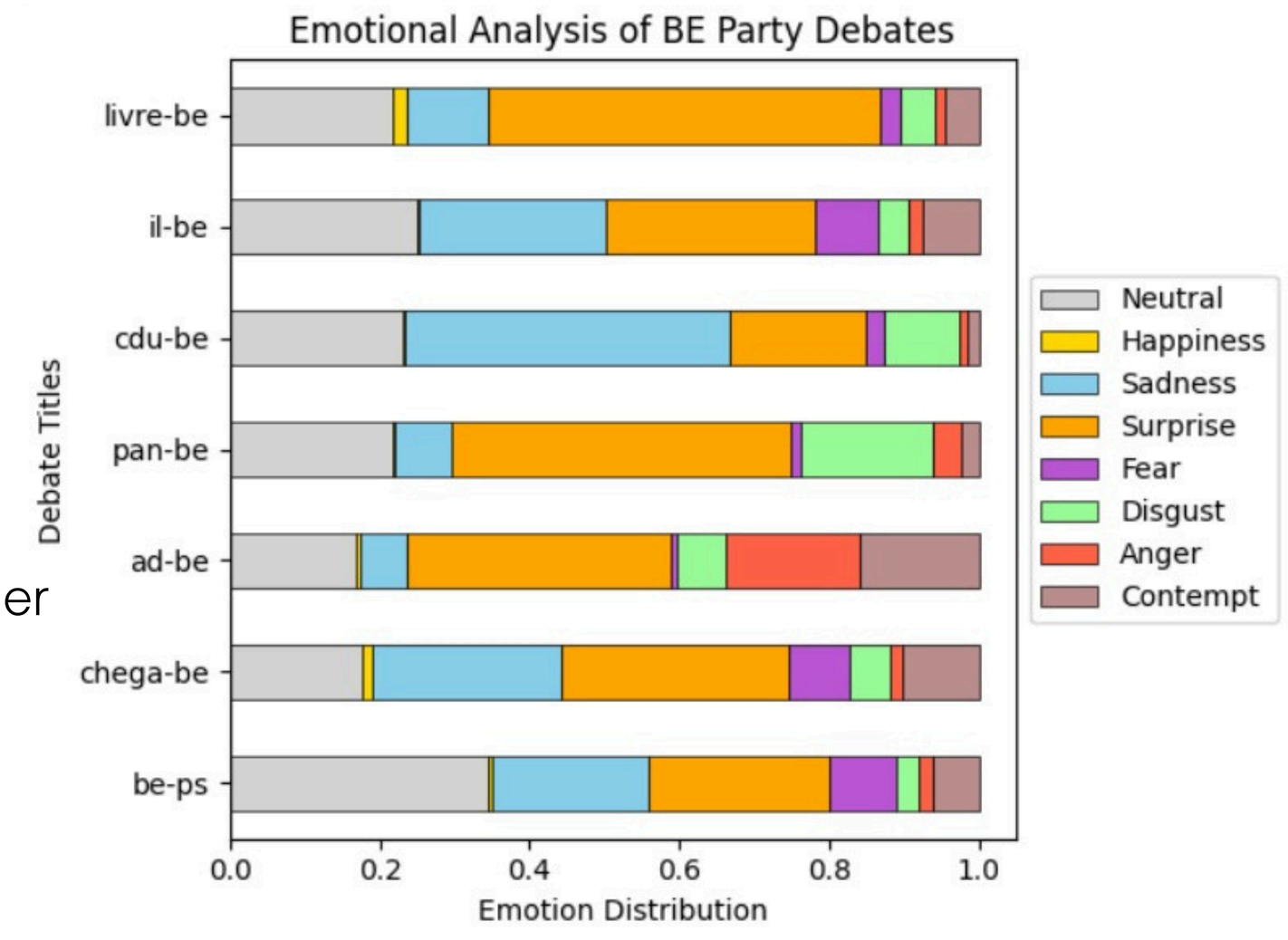


Initial Insights

Emotions

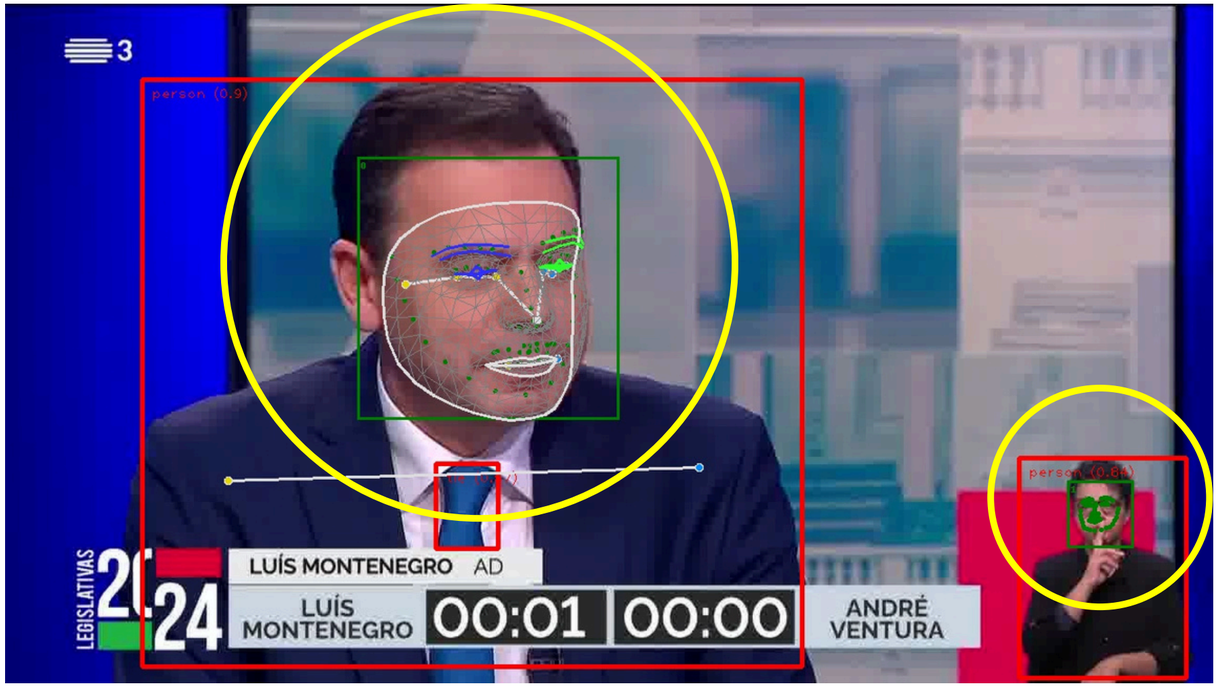
Emotional analysis of **BE debates:**

- Each debate exhibits a distinct emotional profile.
- The **'ad-be'** debate shows a higher frequency of **'anger.'**
- The **'cdu-be**' debate shows a higher frequency of **'sadness.'**



Initial Insights

Is it possible to identify if a 'person' or 'face' in the frame belongs to an interpreter or a politician?



Observations:

- The number of persons in the frame does not reliably indicate whether the individual is a politician or an interpreter.
- The videos vary, with some having no interpreters, others having one, and some having two interpreters.

Importance:

- Distinguishing between politicians and interpreters is crucial for accurate analysis in the first task.

Initial Insights

Main Participants vs. Interpreters

Assumptions

- The main participants (politicians and the moderator) are always positioned at the center of the frame.
- Interpreters consistently appear at the margins of the frame.
- Main participants occupy a larger pixel area within the frame compared to the interpreters.

Hypotheses

- **Position-Based Distinction:** We can distinguish between main participants and interpreters based on their positions in the frame.
- **Area-Based Filtering:** It is also feasible to differentiate these groups by the area they occupy in the frame.

Initial Insights

‘person’ bounding box centers positions for all videos

Hypothesis 1

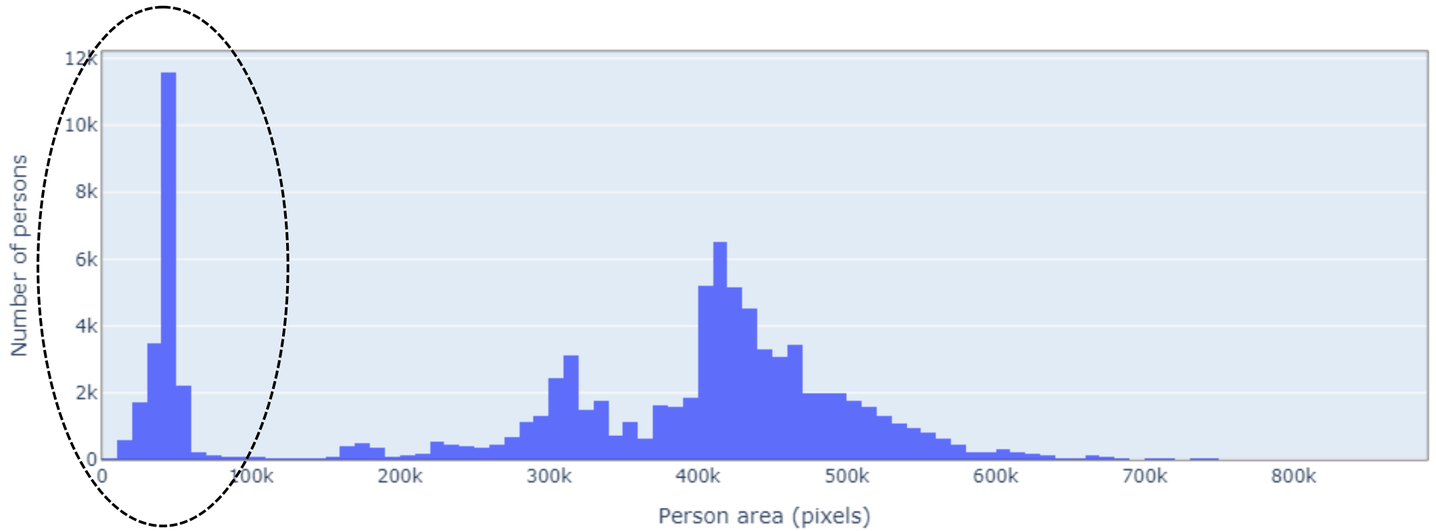
Position-Based Distinction

↓

True

Initial Insights

‘person’ bounding box area distribution for all videos



Hypothesis 2

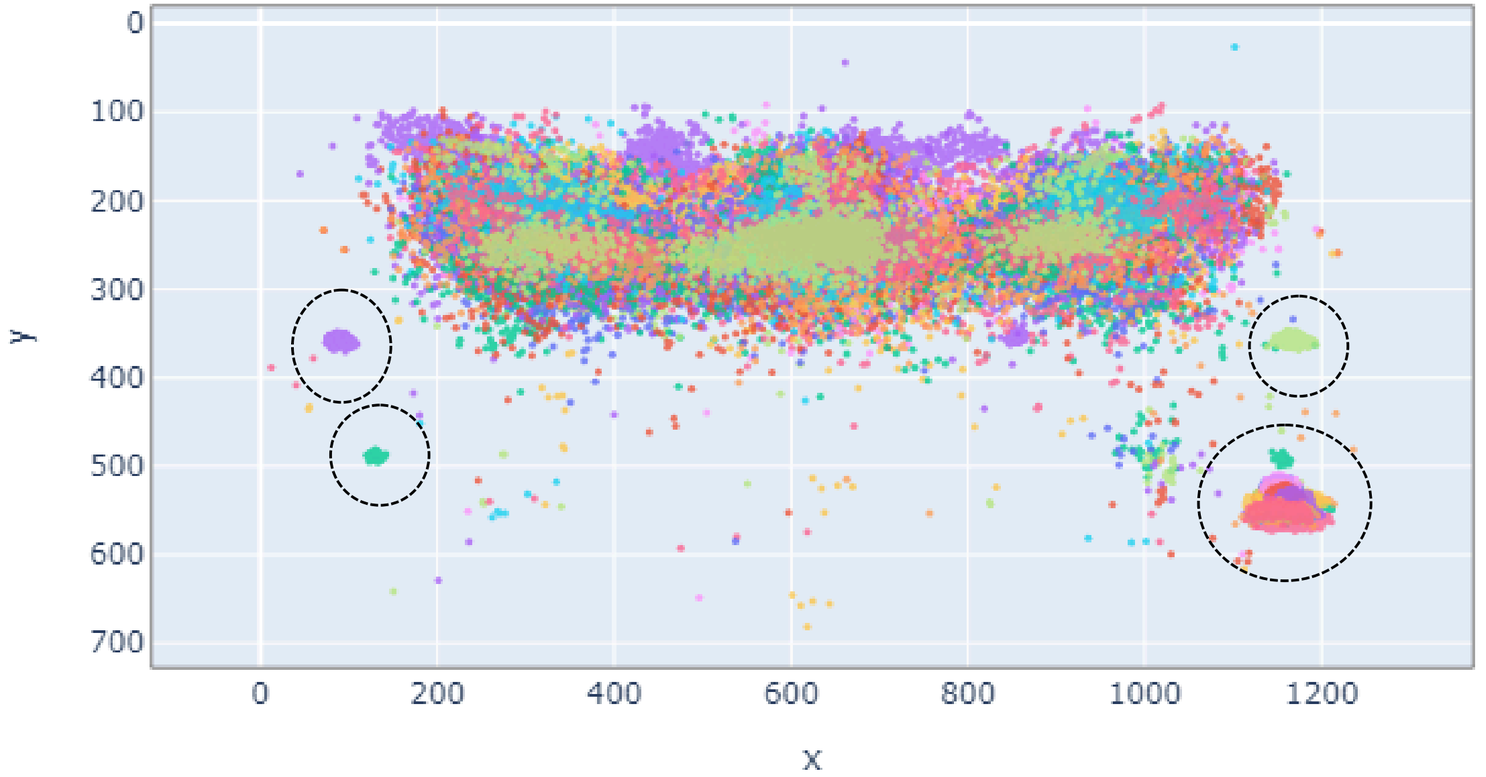
Pixel-Area-Based Filtering

↓

True

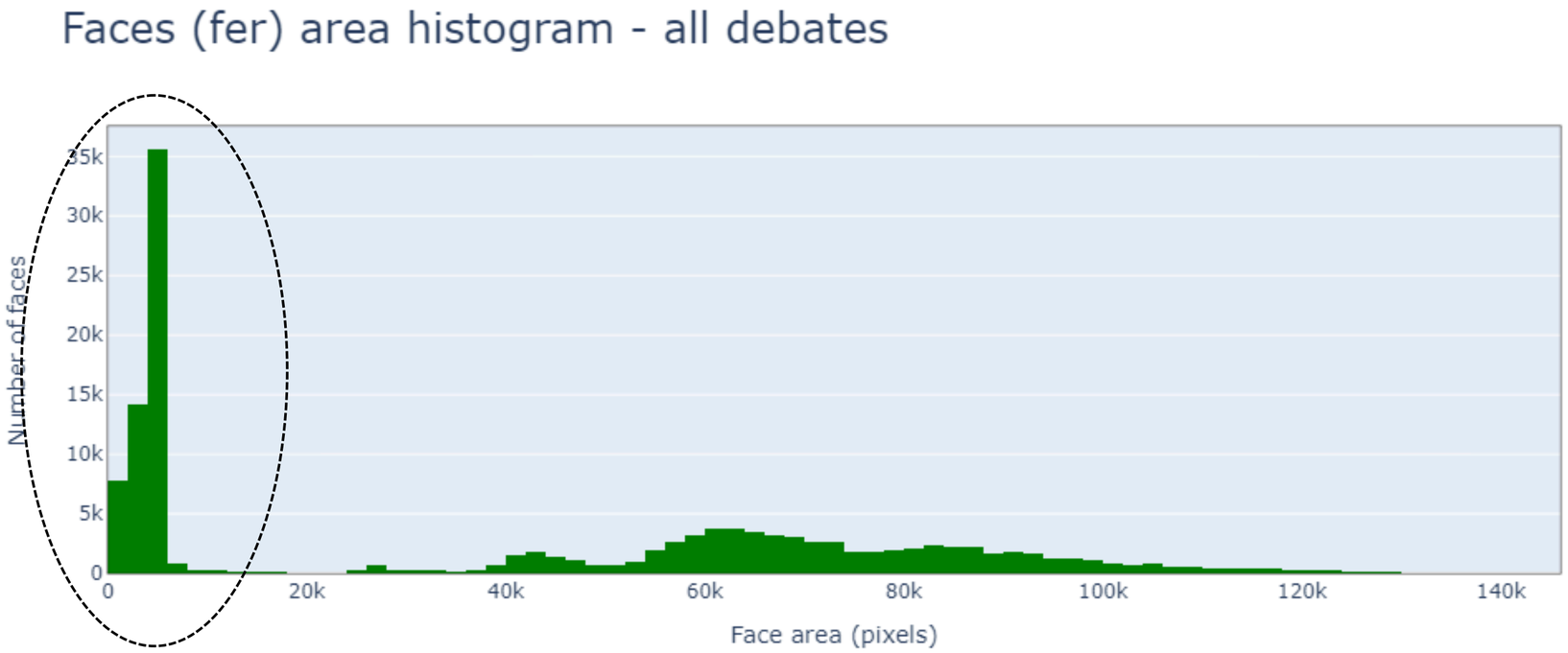
Initial Insights

The same observations are true for the ‘faces’



Initial Insights

The same observations are true for the ‘faces’



Chapter 2 - Dataset Cleaning Process

The cleaning process will comprise two main tasks:

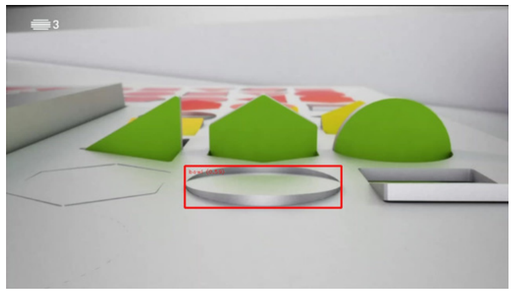
- Removing Non-Debate Frames:
  - Eliminate frames not related to the debate.

- Excluding Sign Language Interpreters:
  - Completely remove the presence of sign language interpreters from the dataset.

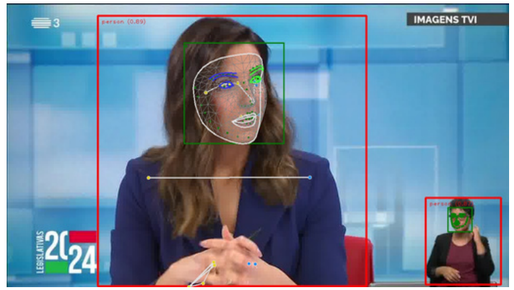
Removing Non-Debate Frames

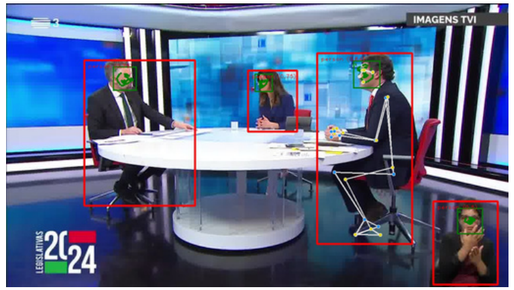
What is our definition of non-debate frames?

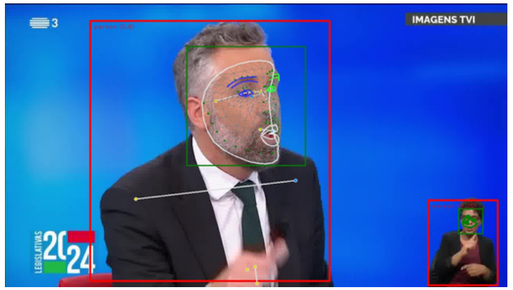
frames we want to eliminate:



frames we want to keep:







Removing Non-Debate Frames

How do we identify them?

By examining the dataframe, we observed that **frames unrelated to the debate** often had **more than 2 empty fields**.

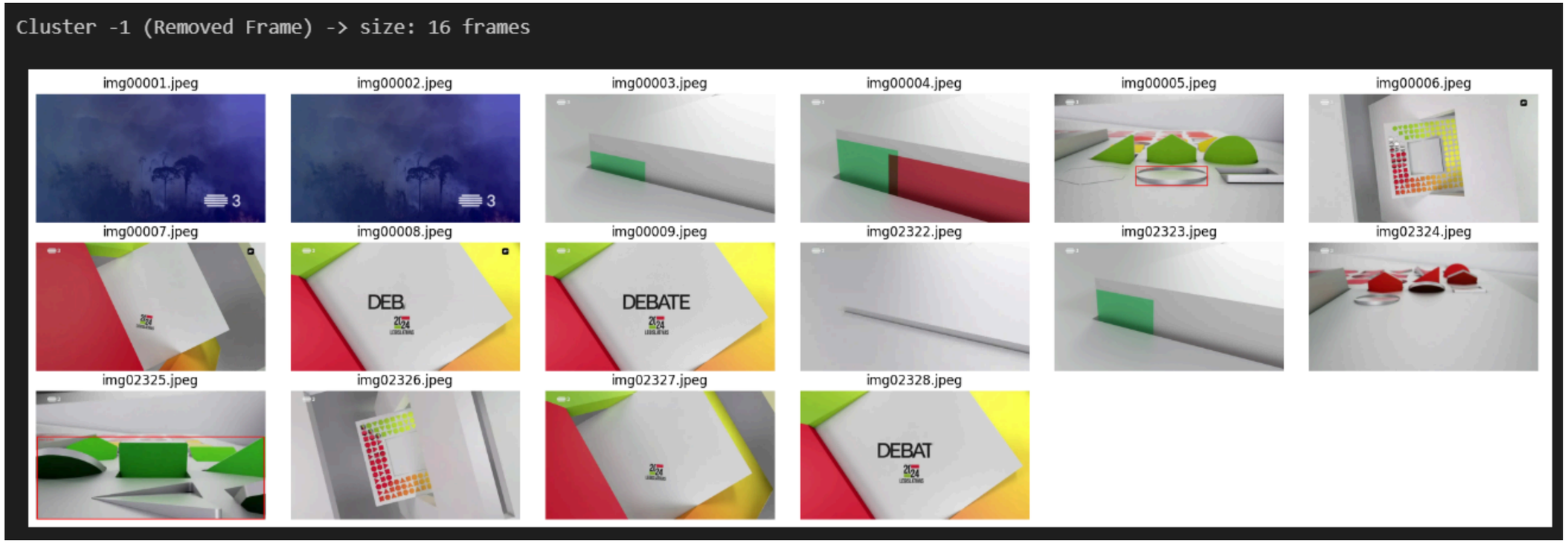
```
Pseudo-code: removed_frames = df[num_empty_fields > 2]
```

Example: ‘chega-ps’ debate results

All frames = 2328 frames
Removed frames = 16 frames
Debate frames = 2312 frames

Removing Non-Debate Frames

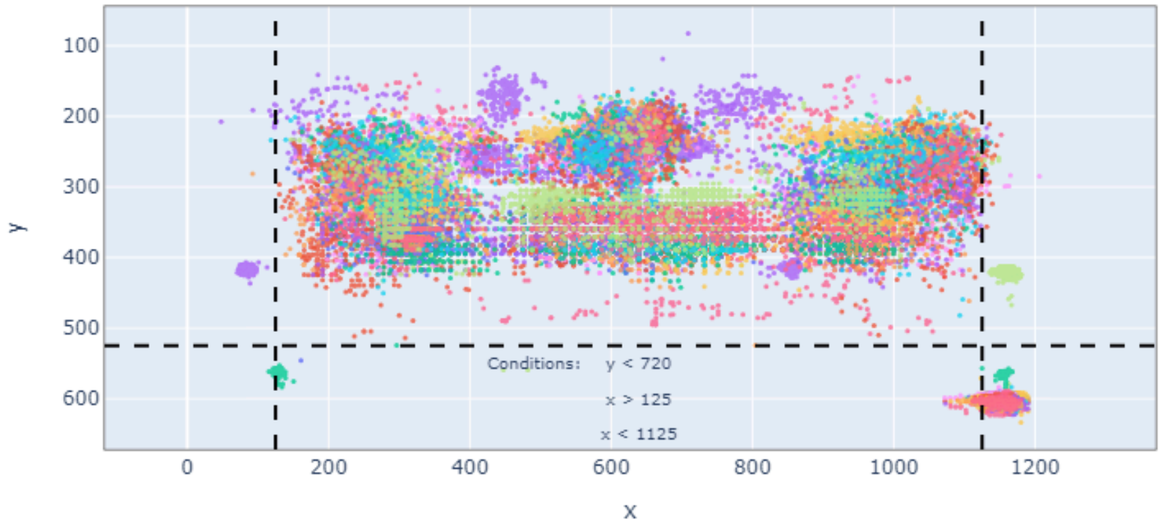
‘chega-ps’ debate results



Excluding Sign Language Interpreters

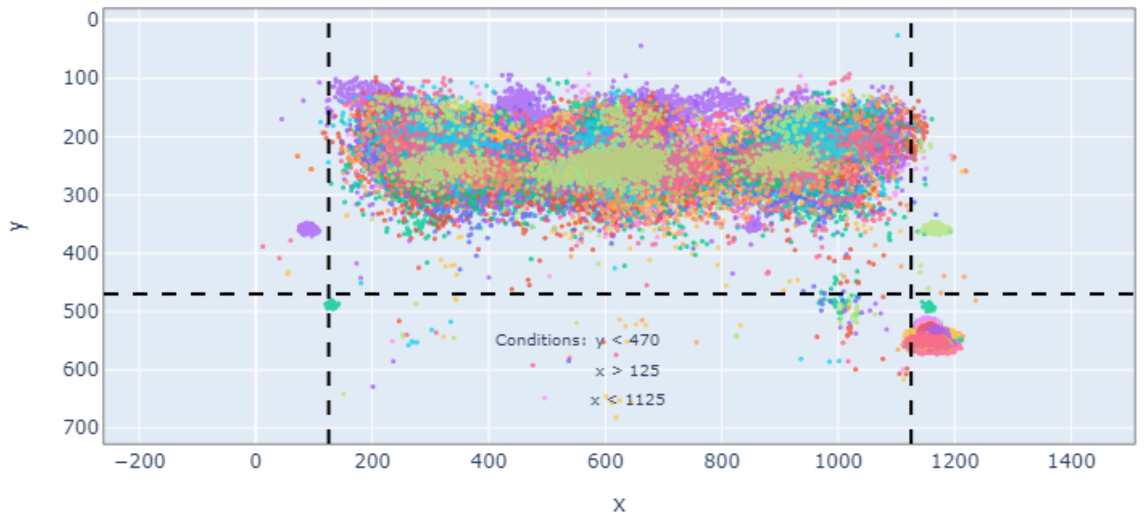
Position-Based Distinction

‘person’s centers




```
for person in persons:
    x = x + w / 2
    y = y + h / 2
    if y < 520 and x > 125 and x < 1125:
      person_in_limits.append(person)
```

‘face’s centers

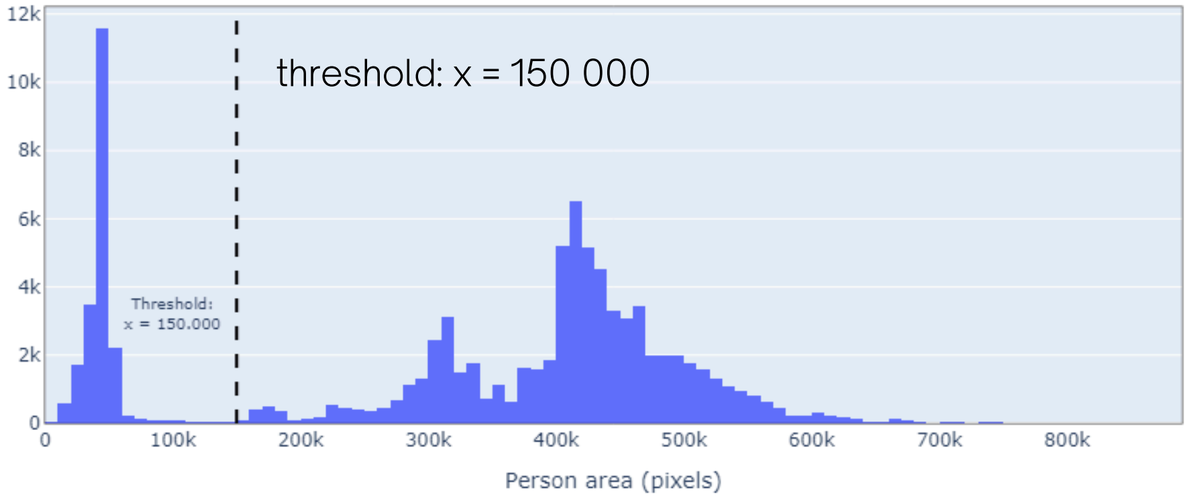



```
for person in persons:
    x = x + w / 2
    y = y + h / 2
    if y < 470 and x > 125 and x < 1125:
      person_in_limits.append(person)
```

Excluding Sign Language Interpreters

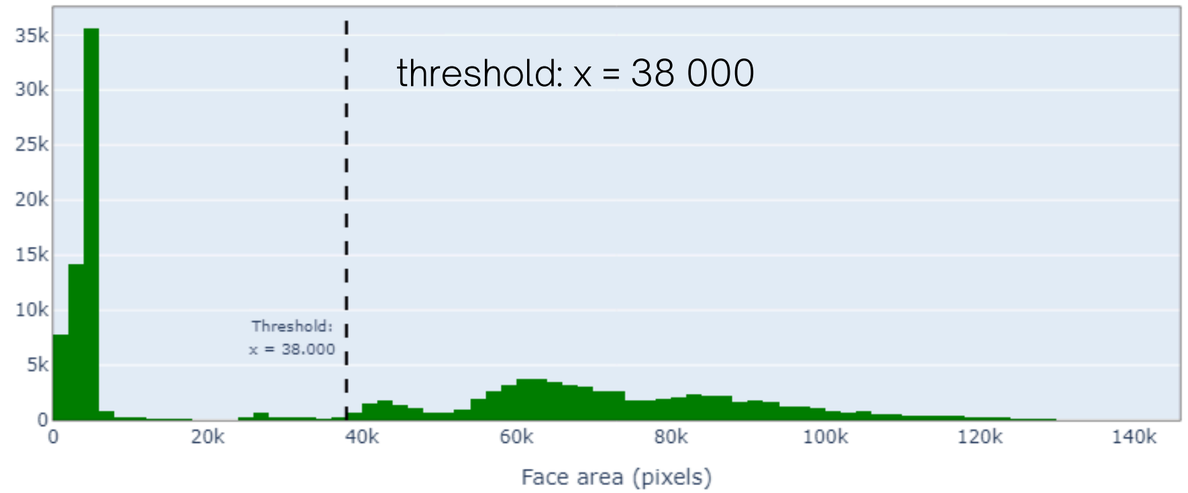
Pixel-Area-Based Filtering

‘person’s area




```
for person in persons:
    area = w * h
    if area > threshold
        valid_person.append(person)
```

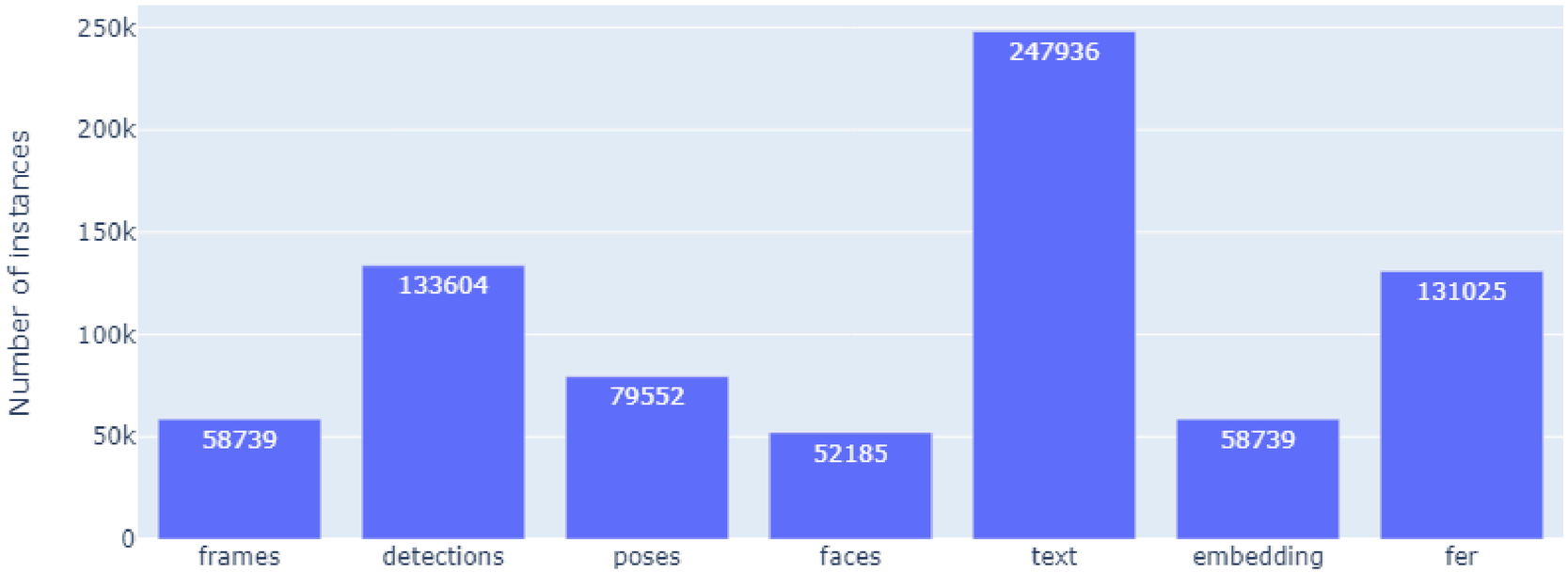
‘face’s area




```
for face in faces:
    area = w * h
    if area > threshold
        valid_faces.append(face)
```

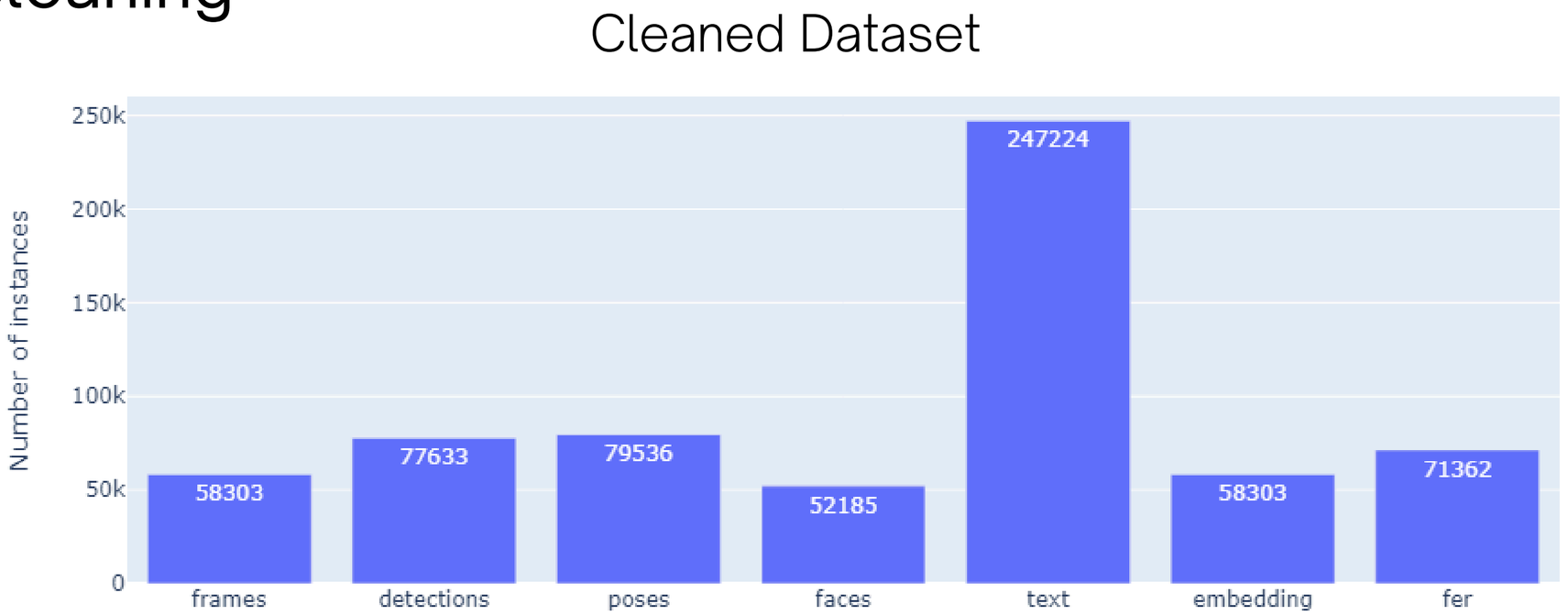
Dataset Cleaning - Results

Before cleaning



Dataset Cleaning - Results

After cleaning



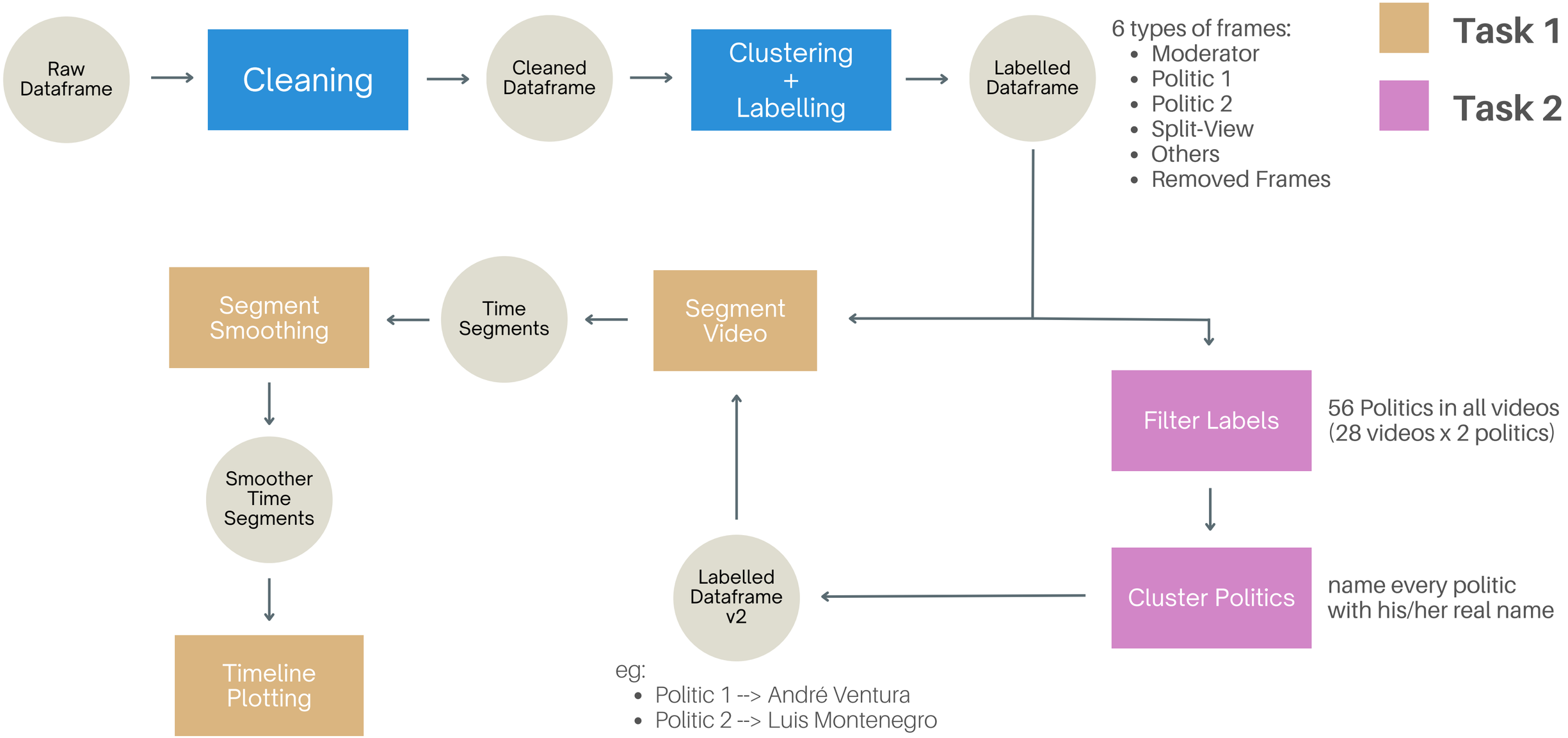
‘frames’
Before: 58 739
After: 58 303
-0,01% reduction

‘detections’
Before: 133 604
After: 77 603
-43% reduction

‘fer’
Before: 131 025
After: 71 362
-46% reduction

Approach overview

Approach overview



Chapter 3 - Task 1

In this chapter, we will discuss the approach and results for the Single Video Analysis task.

Video Segmentation

We will segment the video based on the viewpoint/composition of each frame.

Face Labeling

We will assign a label to each frame/segment in the video.

Task 1 - Aproach

Main objective: Divide the video into segments

A segment is a period of time during which the same viewpoint is continuous

1st Clustering Round

Group Frames into 3 Categories:

- Split-View
- 1-Person-View
- Others

2nd Clustering Round

Subdivide 1-Person-View into:

- Moderator
- Politician 1
- Politician 2

Final Clusters (viewpoints)

- Moderator
- Politician 1
- Politician 2
- Split-View
- Others

1st Clustering Round

For each video, we collected the number of persons ('detections') and faces ('fer') detected on each frame.

Our approach:

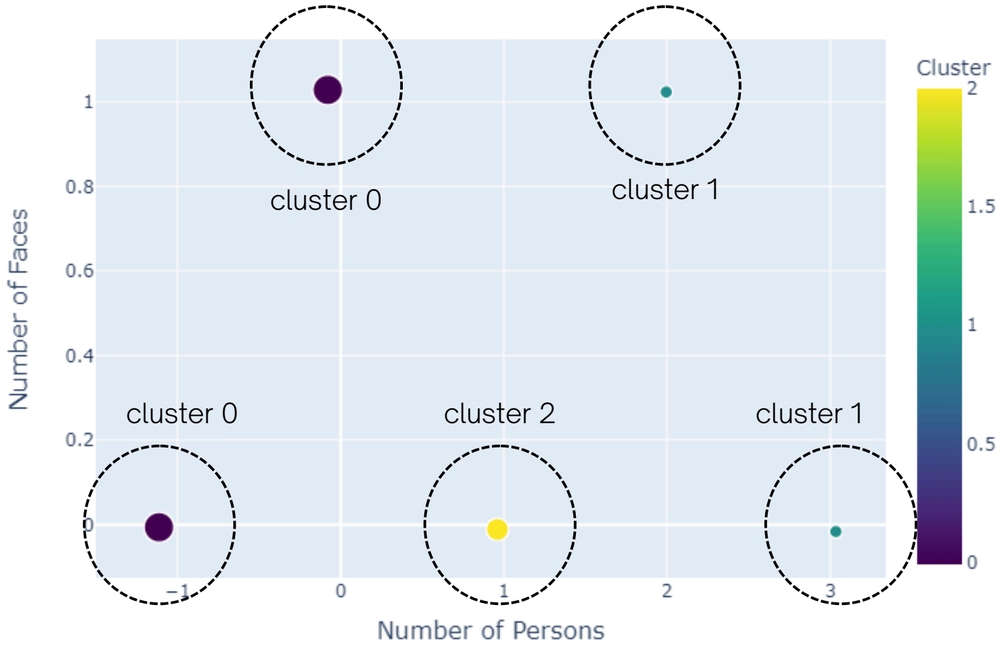
- **Standardized** input data to ensure consistent scaling.
- Utilized Principal Component Analysis **(PCA)** to analyze data variance.
- Employed **Agglomerative Clustering** to group data points based on similarities.

Clustering parameters:

- n_clusters = 3
- metric = ‘euclidean’
- linkage = ‘ward’

1st Clustering Round

The results met our expectations, and we assigned names to each cluster based on the mean number of persons and faces within each cluster.



PCA + Agglomerative Clustering

		Mean persons in clusters	Mean faces in clusters		
Cluster 0 (1327 frames)	→	2.00	1.989	→	**Split-View**
Cluster 1 (253 frames)	→	0.02	0.00	→	**Others**
Cluster 2 (732 frames)	→	1.00	1.00	→	**1-Person-View**

1st Clustering Round

Outputs:

Split-View (1327 frames)	
1-Person-View (732 frames)	
Others (235frames)	

2nd Clustering Round

To subdivide the "1-Person-View" category into clusters that differentiate the appearing person, we used the 1024-dimensional image 'embedding' feature vector.

Our approach:

- **Standardizing Input Data:** Ensured consistent scaling.
- **Applying UMAP:** Used Uniform Manifold Approximation and Projection (UMAP) for dimensionality reduction.
- **Agglomerative Clustering:** Grouped data points based on similarities.

UMAP parameters:

- n_components = 3
- n_neighbors = 15
- min_dist = 0.1

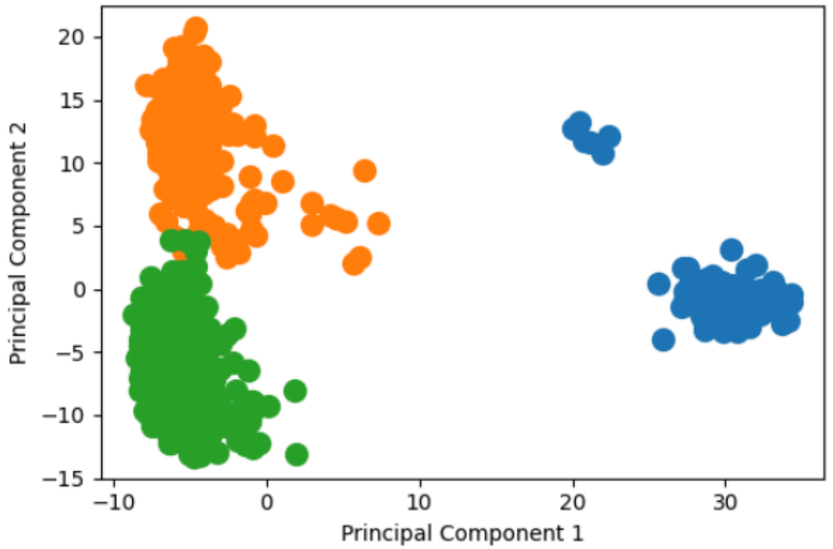
Clustering parameters:

- n_clusters = 3
- metric = ‘euclidean’
- linkage = ‘ward’

Each debate typically features 3 different persons (1 moderator and 2 politicians), except the 'ad-ps' debate which includes 3 moderators.
Set n_clusters to 3 based on this assumption.

2nd Clustering Round

The results met our expectations, and we assigned names to each cluster based on the dominant individual identified within each cluster.



UMAP + Agglomerative Clustering

1-Person-View (738 frames) →

Cluster 0 (405 frames) → **Politician 1** (405 frames)

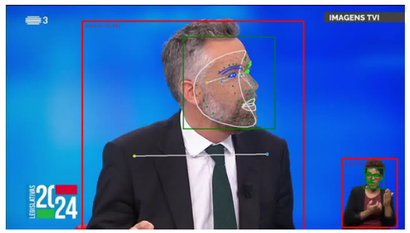
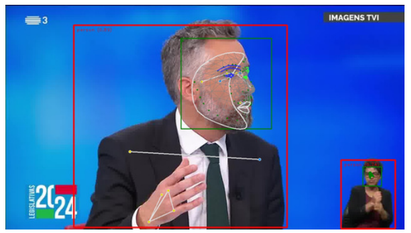
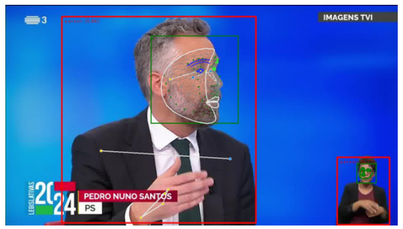
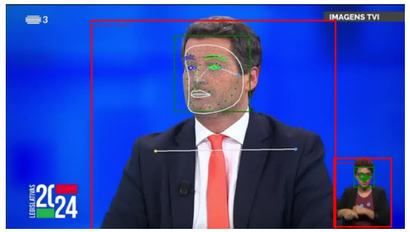
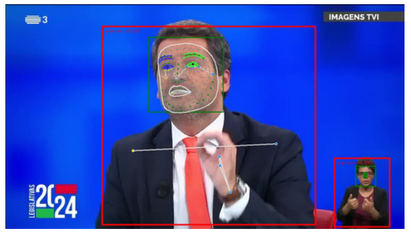
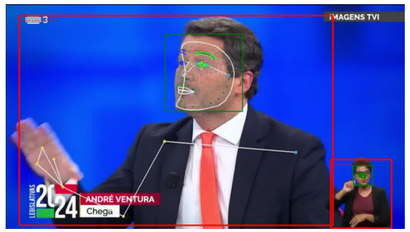
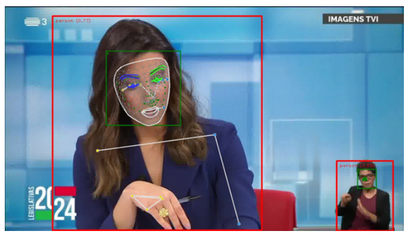
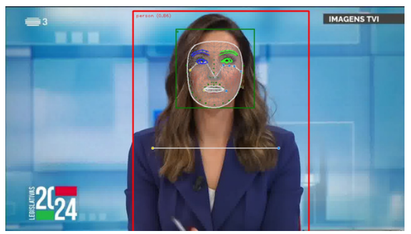
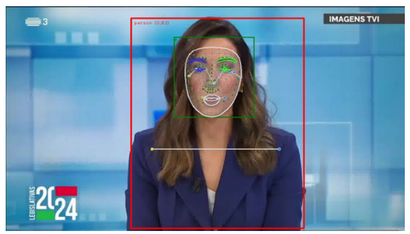
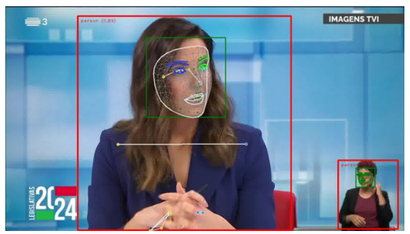
Cluster 1 (222 frames) → **Politician 2** (222 frames)

Cluster 2 (111 frames) → **Moderator** (111 frames)

Assumption: The cluster with less frames is assigned to Moderator

2nd Clustering Round

Outputs:

Politician 1 (405 frames)				
Politician 2 (222 frames)				
Moderator (111 frames)				

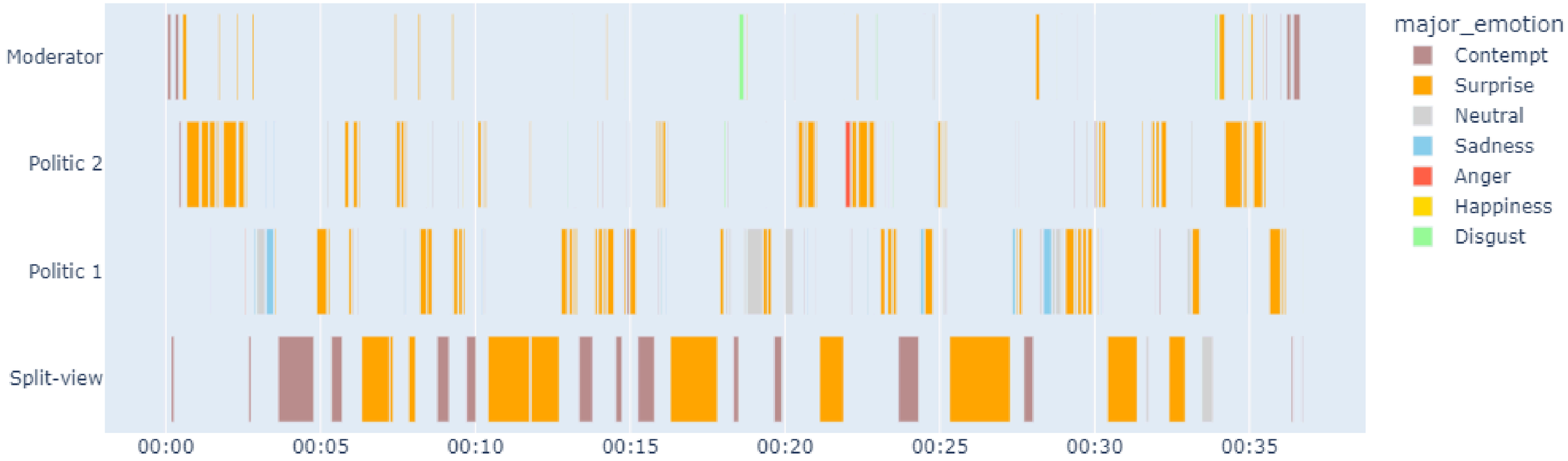
Labelled dataframe

Code output:

```
Cluster -1 (Removed Frame) -> size: 16 frames
Cluster 0 (Others) -> size: 261 frames
Cluster 1 (Split-view) -> size: 1313 frames
Cluster 2 (Politic 1) -> size: 405 frames
Cluster 3 (Politic 2) -> size: 222 frames
Cluster 4 (Moderator) -> size: 111 frames
```

Video Segmentation

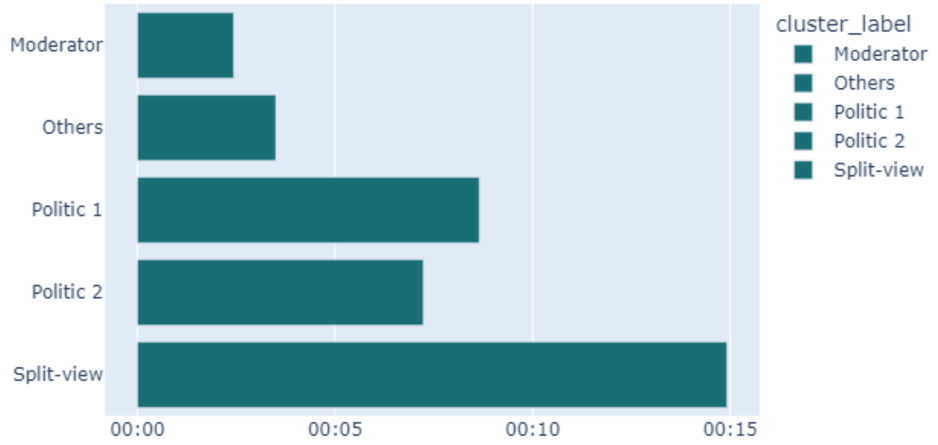
The time segments of each viewpoint of the video with the major emotion present in the corresponding time interval



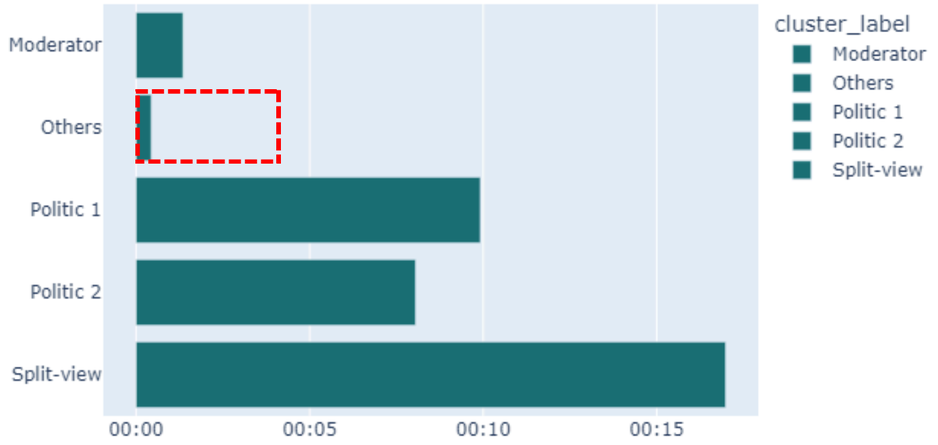
Segment Smoothing

In order to remove noise, we applied moving average as a smoothing technique to improve the quality of our video segmentation

Raw Timeline



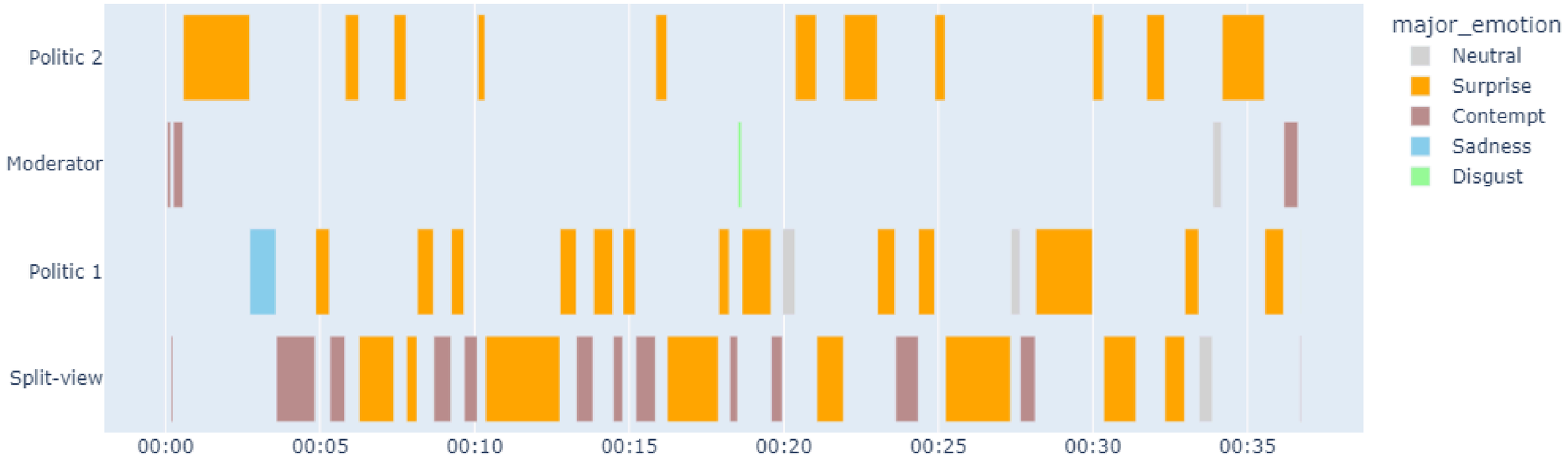
Smoothed Timeline



We can notice that after the smoothing process declares that most of the ‘others’ viewpoint frames are considered as noise, since these kind of time intervals are pretty short in comparison to the remaining

Video Timeline (after smoothing)

This is the final timeline for the ‘chega-be’ video after the smoothing process:



Chapter 4 - Task 2

Multi Video Analysis

Renaming the Labels

We will replace, unsupervised, the 'Politician 1' and 'Politician 2' labels with their corresponding names.

Clustering Politicians

In order to decide which name to assign to each politic in a video we did the following:

Gather all debates from a specific party (7 videos per party) → Extract Politicians (7*2 = 14 Politicians) → Cluster Politicians

Expected results:

- 1 big cluster comprising 7 politicians (same person)
- 7 small clusters each for a different politician

Clustering Politicians

How?

For each ‘politician’ we extracted their **200 best faces** to describe its cluster;

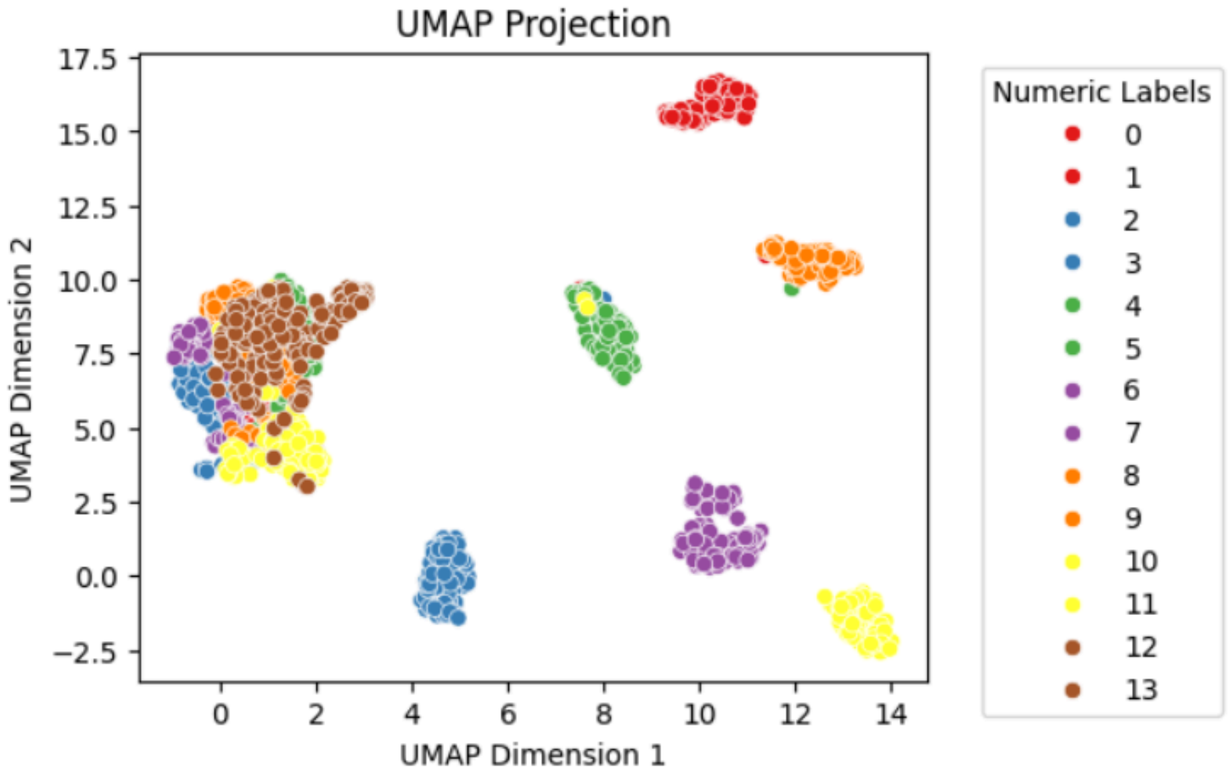
- The ‘best faces’ are those where the politician is in a frontal-view, so the embedding is more precise;
- We used the ‘poses’ column to filter this out.

Our approach:

- **Feature vector:** 2800 face embeddings (14*200)
- **Supervised UMAP:** We labeled each face with their corresponding debate/politician origin.
- **Agglomerative Clustering:** Grouped data points based on similarities.

Clustering Politicians

Results for ‘chega’ party:



The biggest cluster is the person representing the party.

We can now name all the corresponding politicians with the name “André Ventura”

Final Results

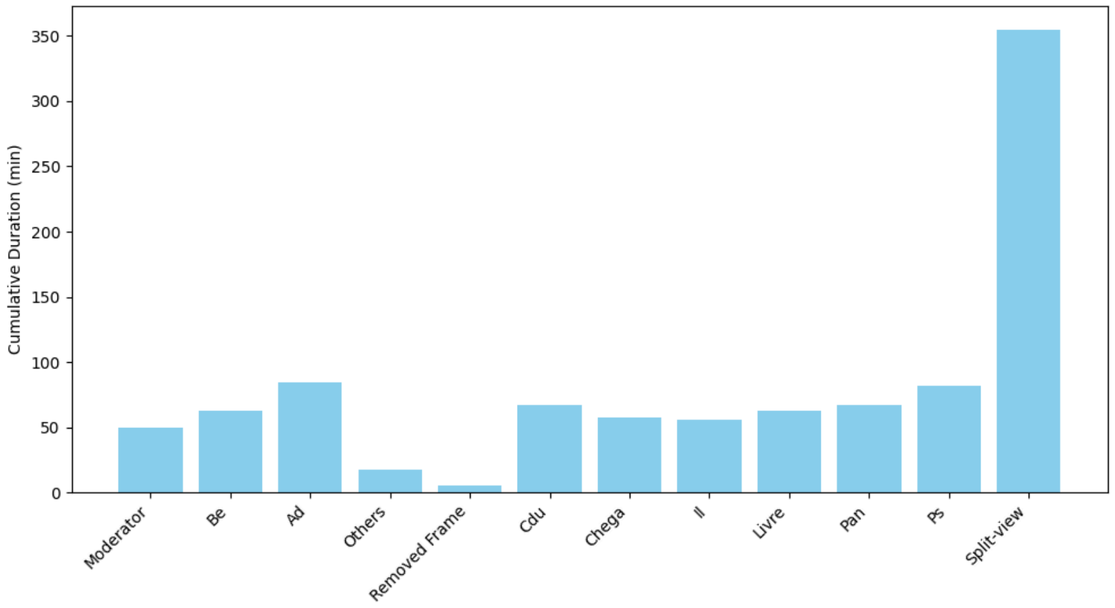
Results for all parties

Debate	Politic 1	Politic 2	Debate	Politic 1	Politic 2	Debate	Politic 1	Politic 2	Debate	Politic 1	Politic 2
ad-be	AD	BE	cdu-be	BE	CDU	il-ps	PS	IL	pan-chega	CHEGA	PAN
ad-cdu			cdu-ps	PS	CDU	livre-be	LIVRE	BE	pan-il	IL	PAN
ad-chega	AD	CHEGA	chega-be	CHEGA	BE	livre-cdu			pan-livre		
ad-il	IL	AD	chega-cdu	CHEGA	CDU	livre-chega	CHEGA	LIVRE	pan-ps	PAN	PS
ad-livre	LIVRE	AD	chega-il	CHEGA	IL	livre-il	IL	LIVRE			
ad-pan	PAN	AD	chega-ps	CHEGA	PS	livre-ps					
ad-ps	PS	AD	il-be	IL	BE	pan-be	BE	PAN			
be-ps	PS	BE	il-cdu	IL	CDU	pan-cdu	PAN	CDU			

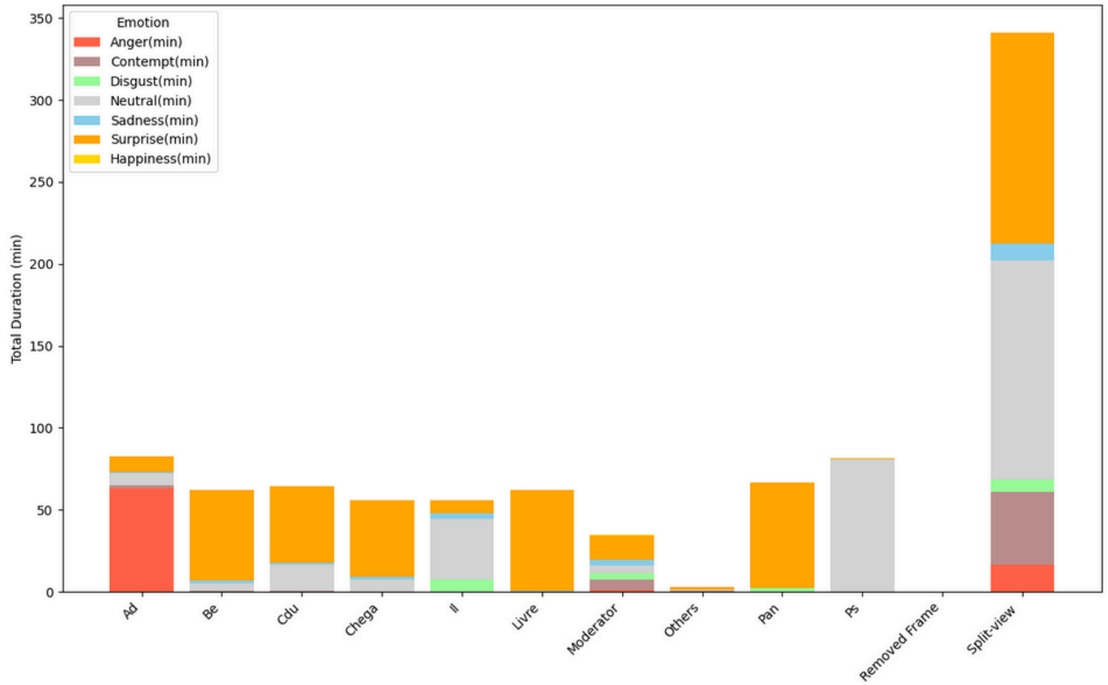
Final Results

Overall summary of 2024 election debates

Cumulative duration for each viewpoint



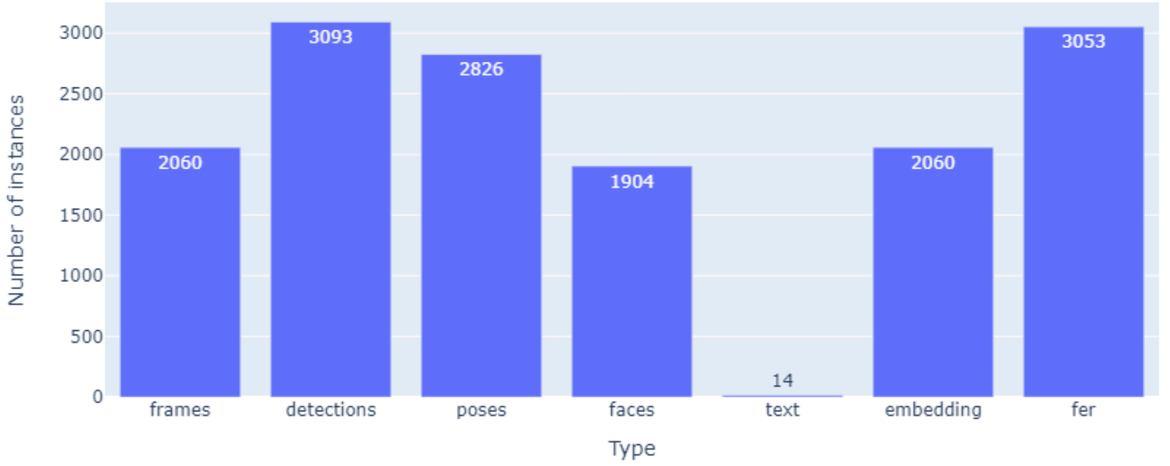
Total duration of each emotion for each viewpoint



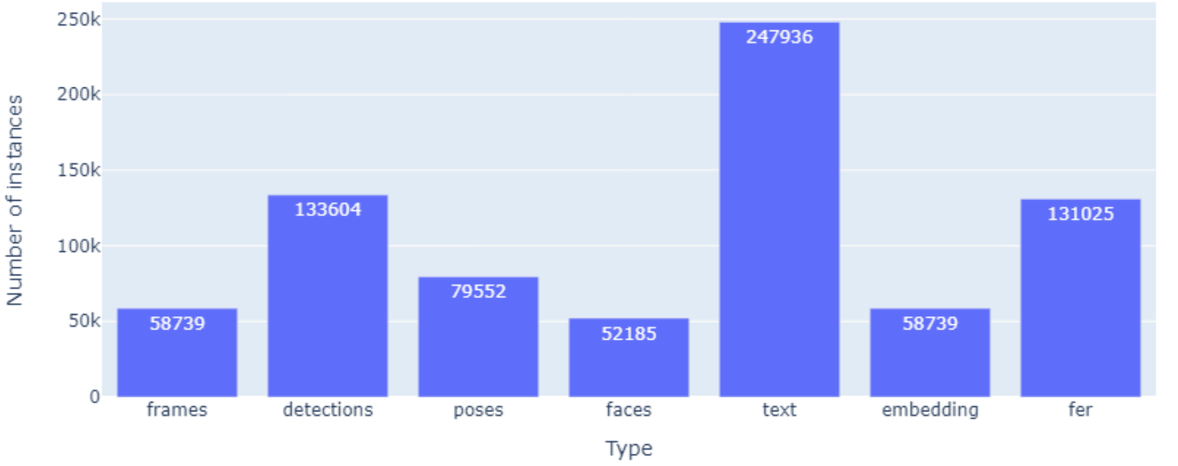
These results were achieved by naming the clusters manually

Plots

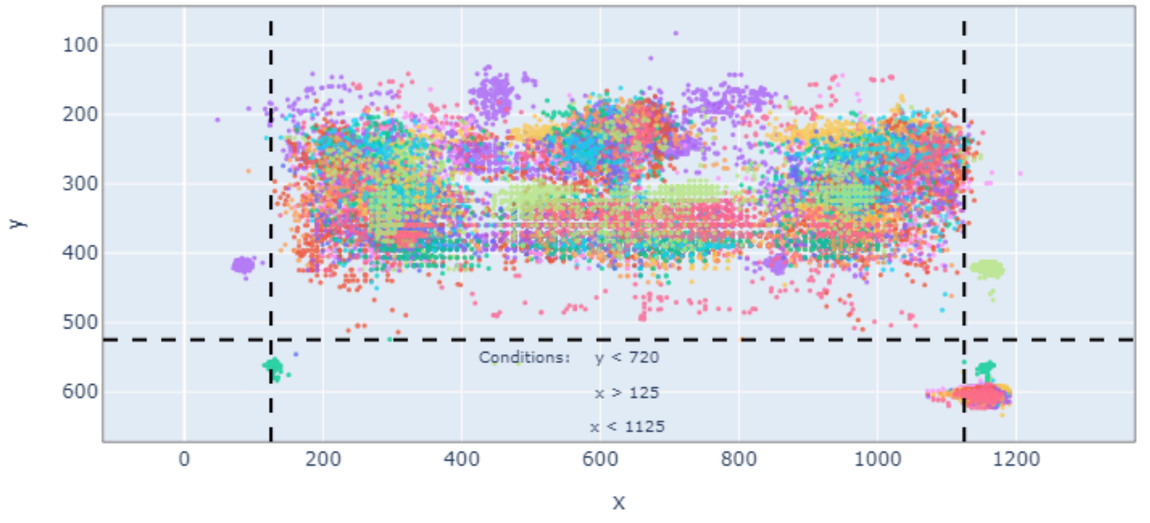
Dataframe Characterization of ad-be video



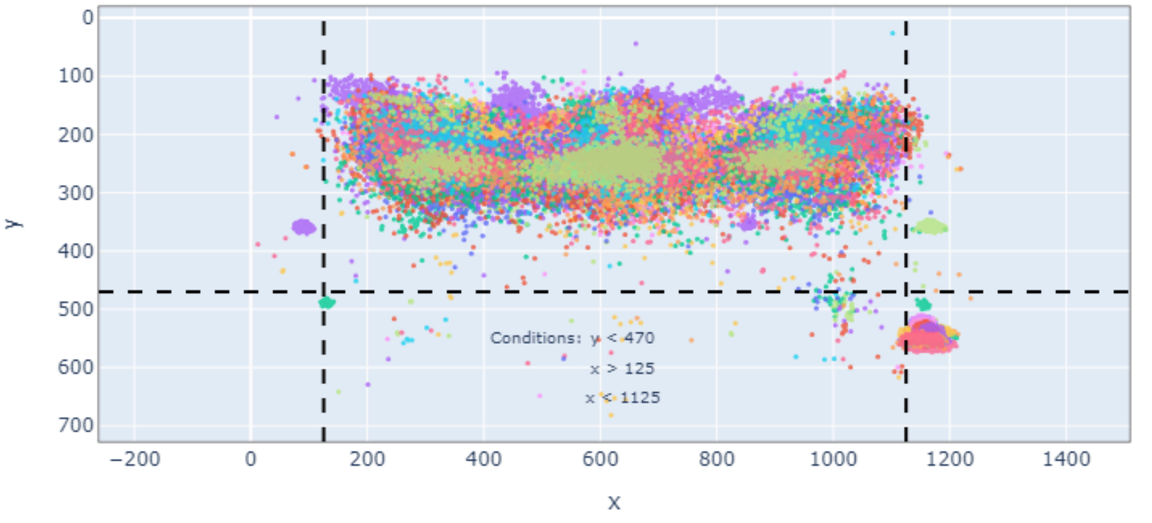
Dataframe Characterization of all videos



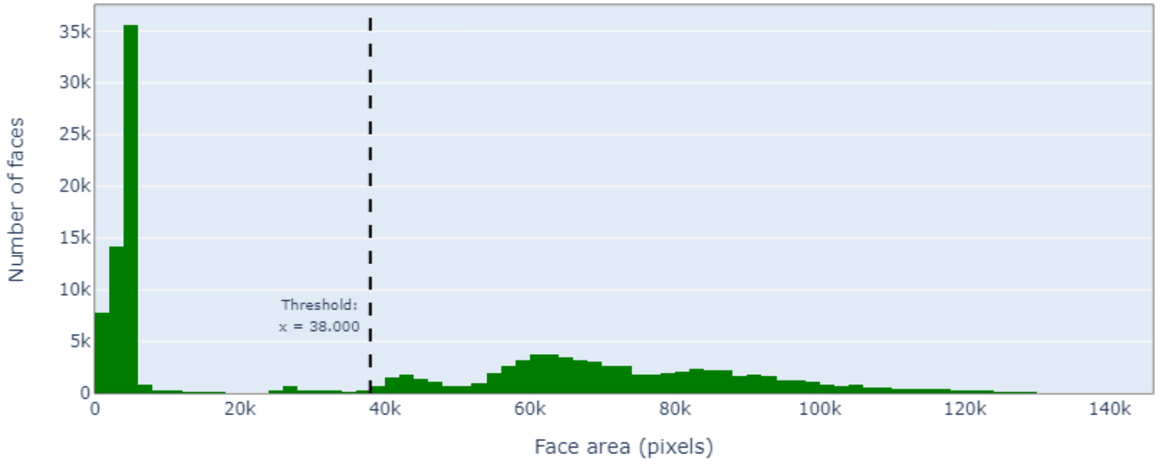
Person centers - all debates



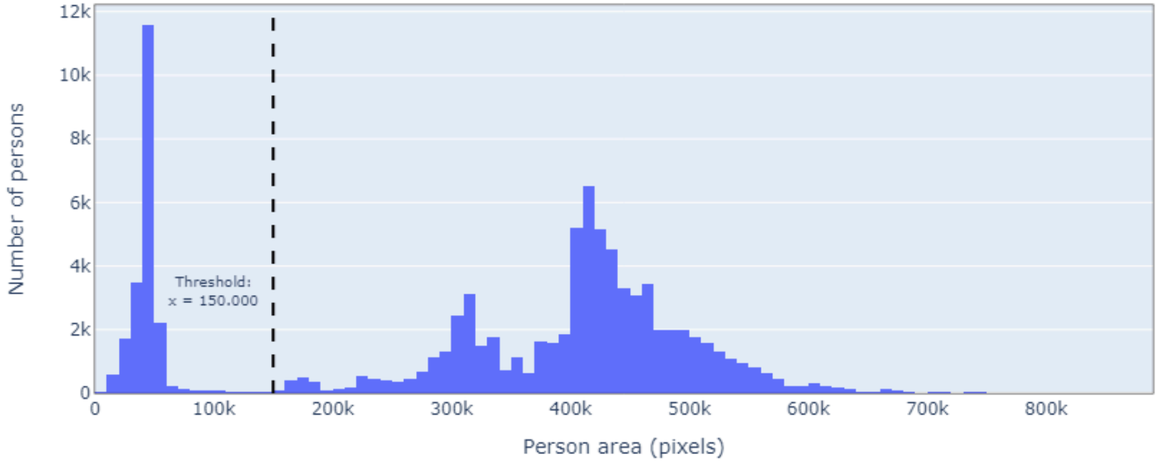
Face centers - all debates



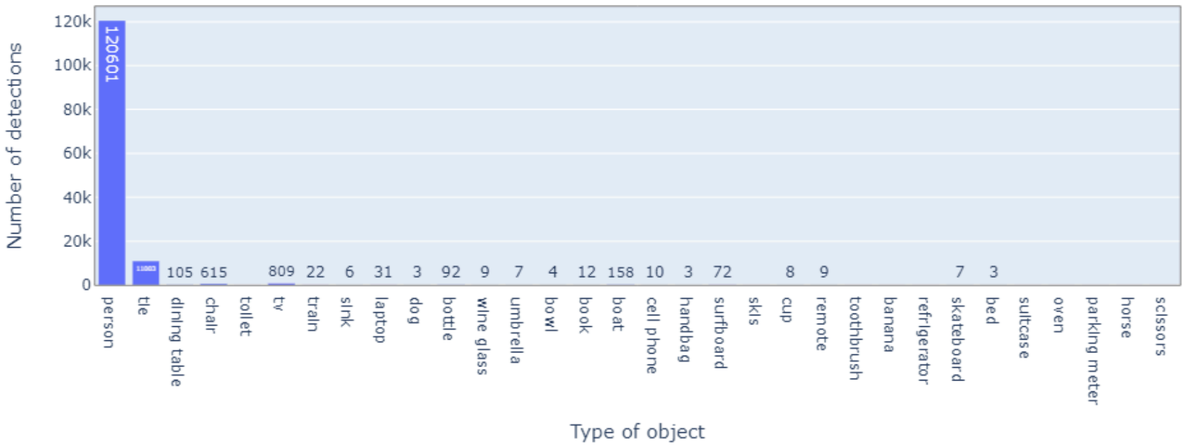
Faces (fer) area histogram - all debates



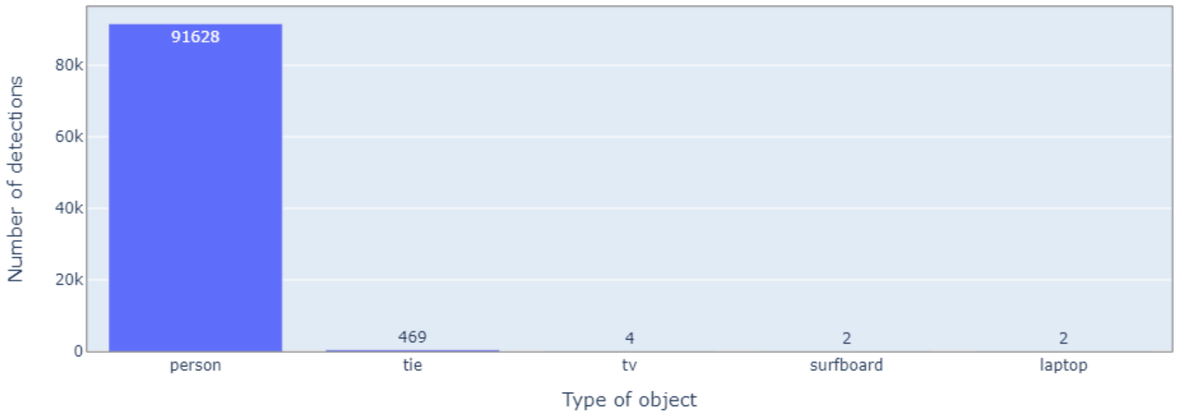
Person (detections) area histogram - all debates



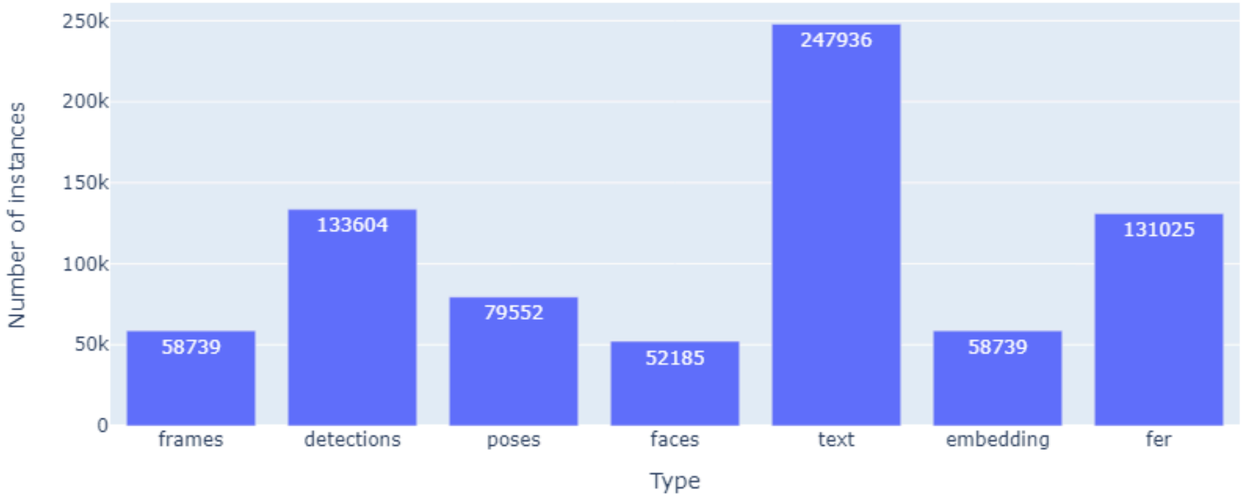
Objects detected with confidence higher than 0% - all debates



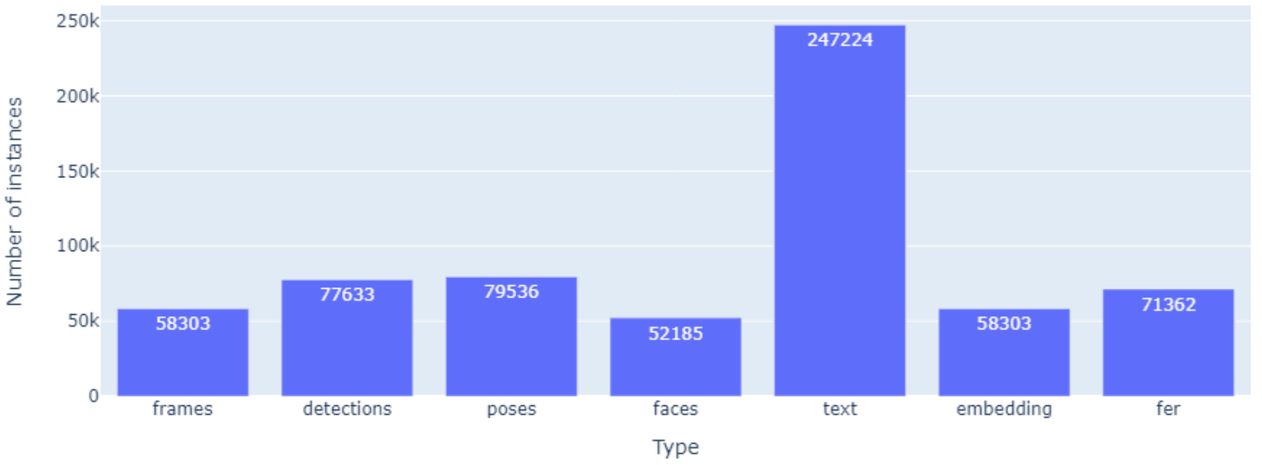
Objects detected with confidence higher than 75% - all debates

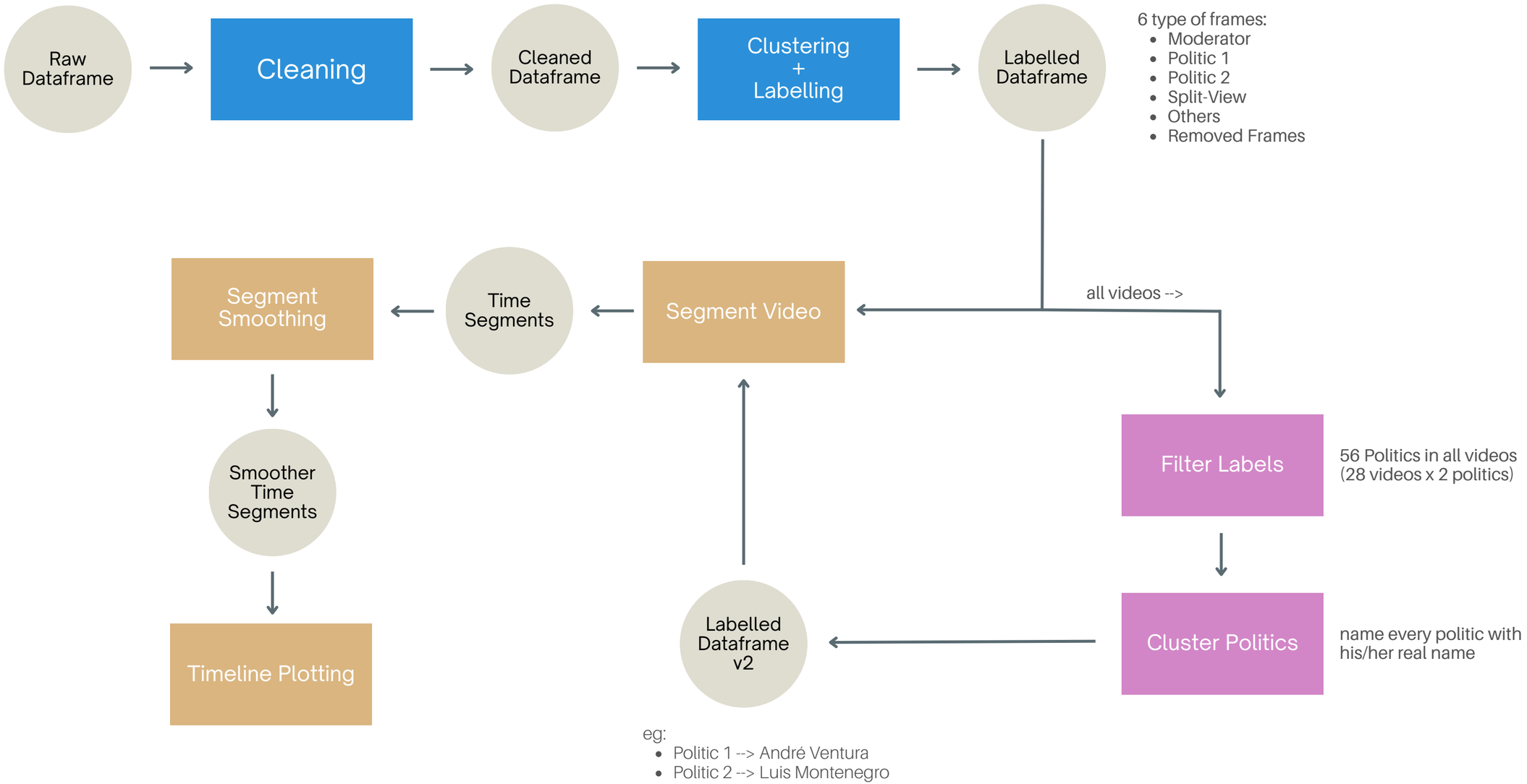


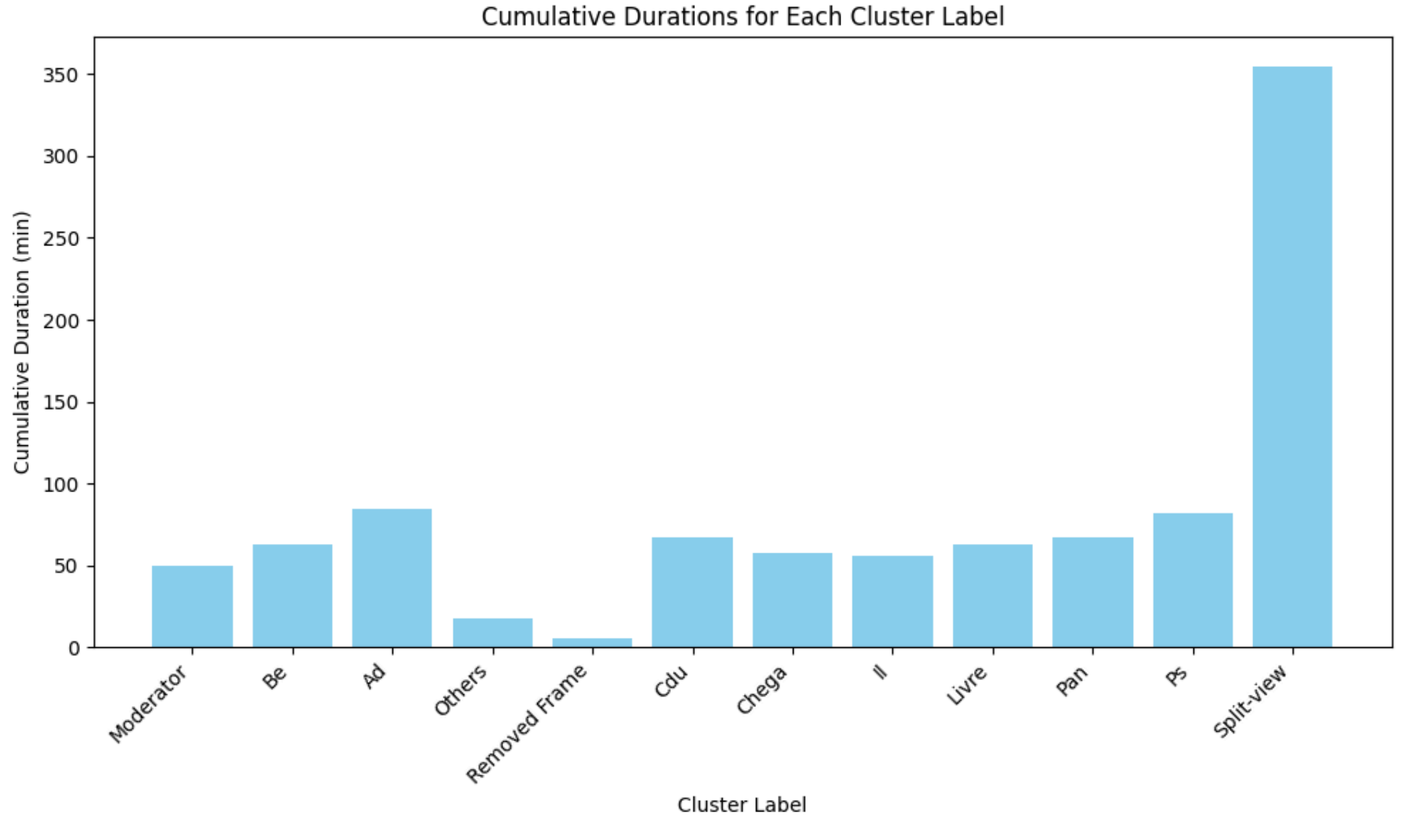
Dataframe Characterization of all videos

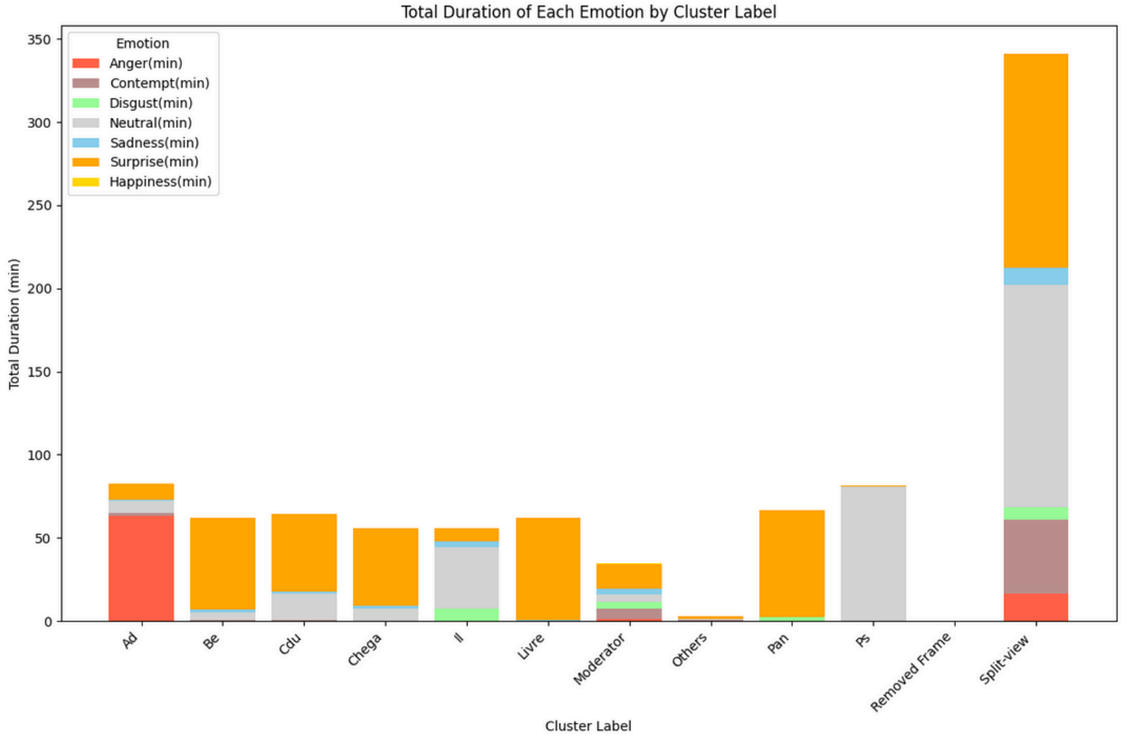


Dataframe Characterization of all videos after cleaning

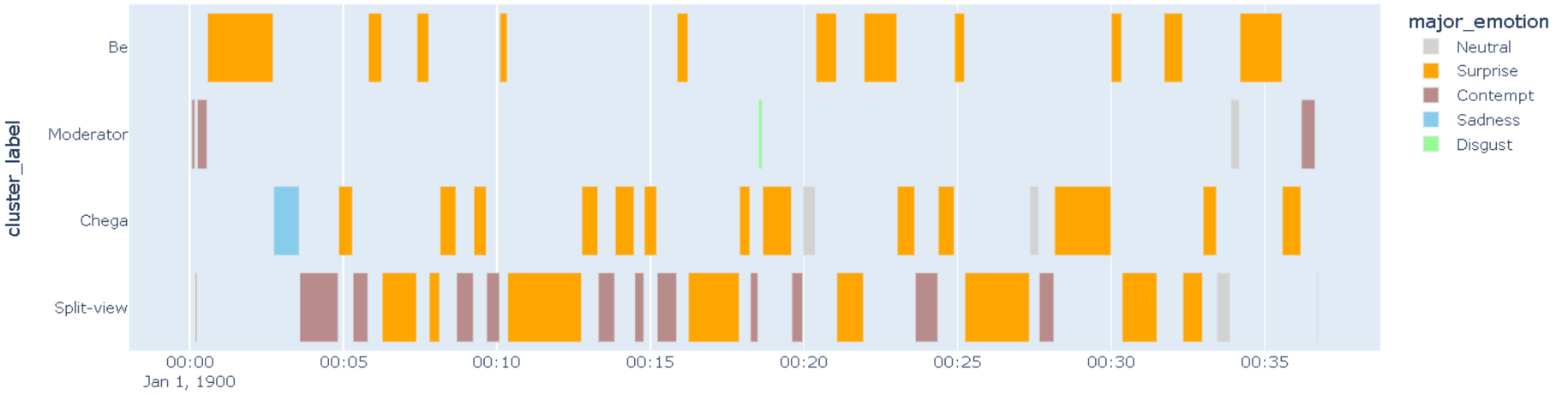



Raw Dataframe
Cleaning
Cleaned Dataframe
Clustering + Labelling
Labelled Dataframe
6 type of frames:
Moderator
Politic 1
Politic 2
Split-View
Others
Removed Frames
all videos -->
Segment Video
Time Segments
Segment Smoothing
Smoother Time Segments
Timeline Plotting
Filter Labels
56 Politics in all videos (28 videos x 2 politics)
Cluster Politics
name every politic with his/her real name
Labelled Dataframe v2
eg:
Politic 1 --> André Ventura
Politic 2 --> Luis Montenegro

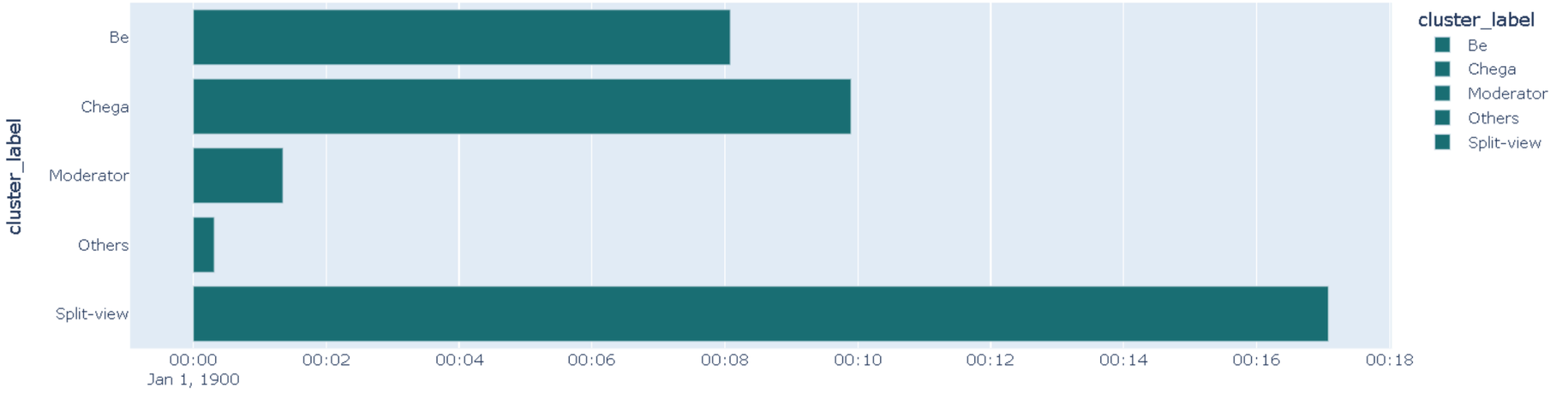

Cumulative Durations for Each Cluster Label
Cumulative Duration (min)
0
50
100
150
200
250
300
350
Moderator
Be
Ad
Others
Removed Frame
Cdu
Chega
Il
Livre
Pan
Ps
Split-view
Cluster Label


Total Duration of Each Emotion by Cluster Label
Emotion
Anger(min)
Contempt(min)
Disgust(min)
Neutral(min)
Sadness(min)
Surprise(min)
Happiness(min)
Total Duration (min)
0
50
100
150
200
250
300
350
Ad
Be
Cdu
Chega
Il
Livre
Moderator
Others
Pan
Ps
Removed Frame
Split-view
Cluster Label

Video Timeline - chega-be

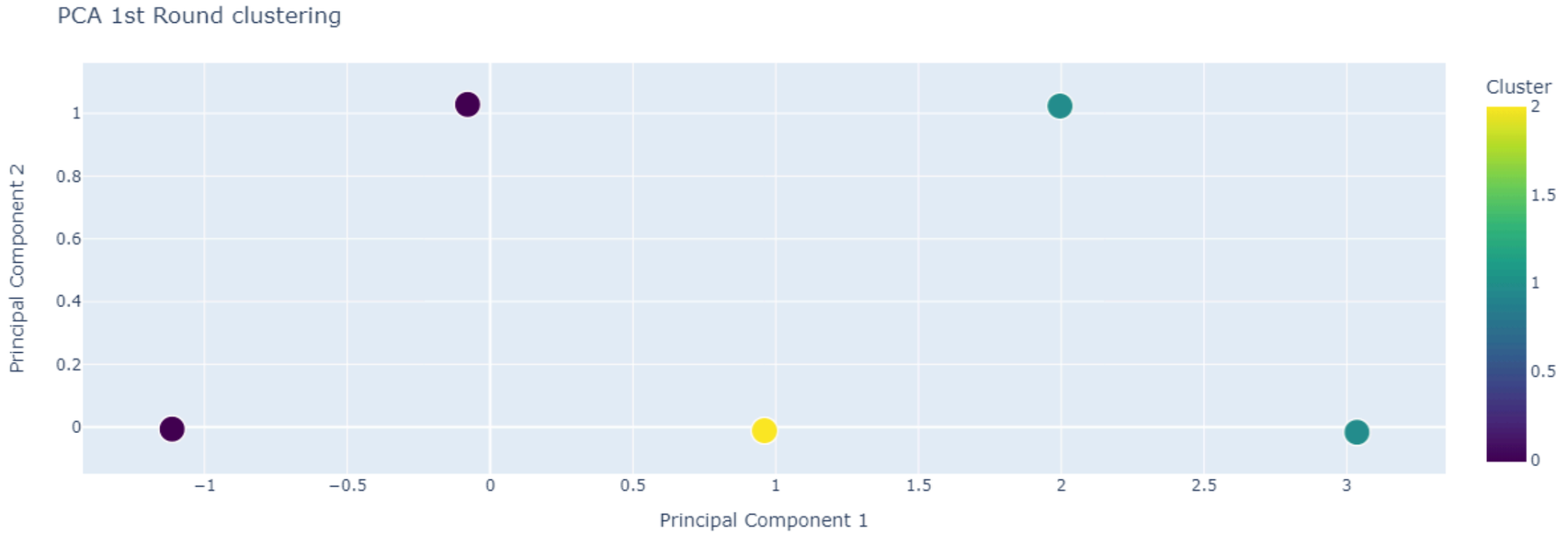


Screen Time per Cluster - chega-be



	cluster_label	Contempt	Disgust	Neutral	Sadness	Surprise	duration
0	Be	0.0	0.0	0.0	0.0	484.0	485
1	Chega	0.0	0.0	43.0	51.0	500.0	594
2	Moderator	55.0	8.0	18.0	0.0	0.0	81
3	Split-view	376.0	0.0	27.0	0.0	622.0	1025

	cluster_label	duration	duration(min)
0	Be	485	08:05
1	Chega	594	09:54
2	Moderator	81	01:21
3	Others	19	00:19
4	Split-view	1025	17:05


PCA 1st Round clustering
Principal Component 2
1
0.8
0.6
0.4
0.2
0
−1
−0.5
0
0.5
1
1.5
2
2.5
3
Principal Component 1
Cluster
2
1.5
1
0.5
0