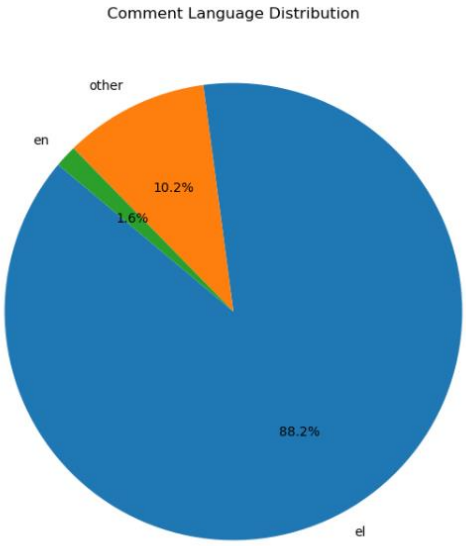
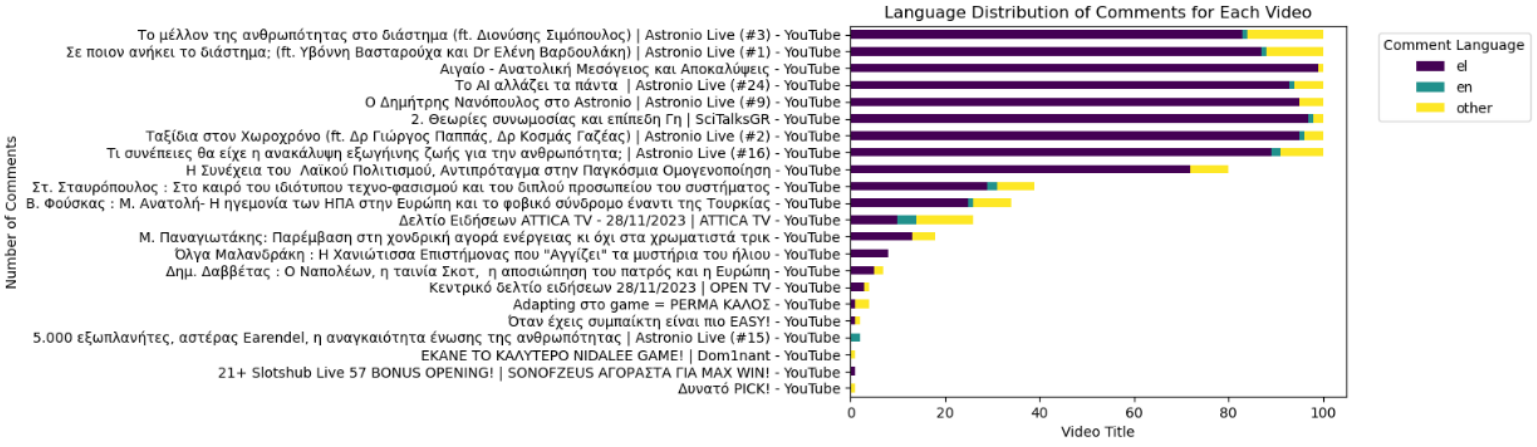
Report

III Improve language detection.

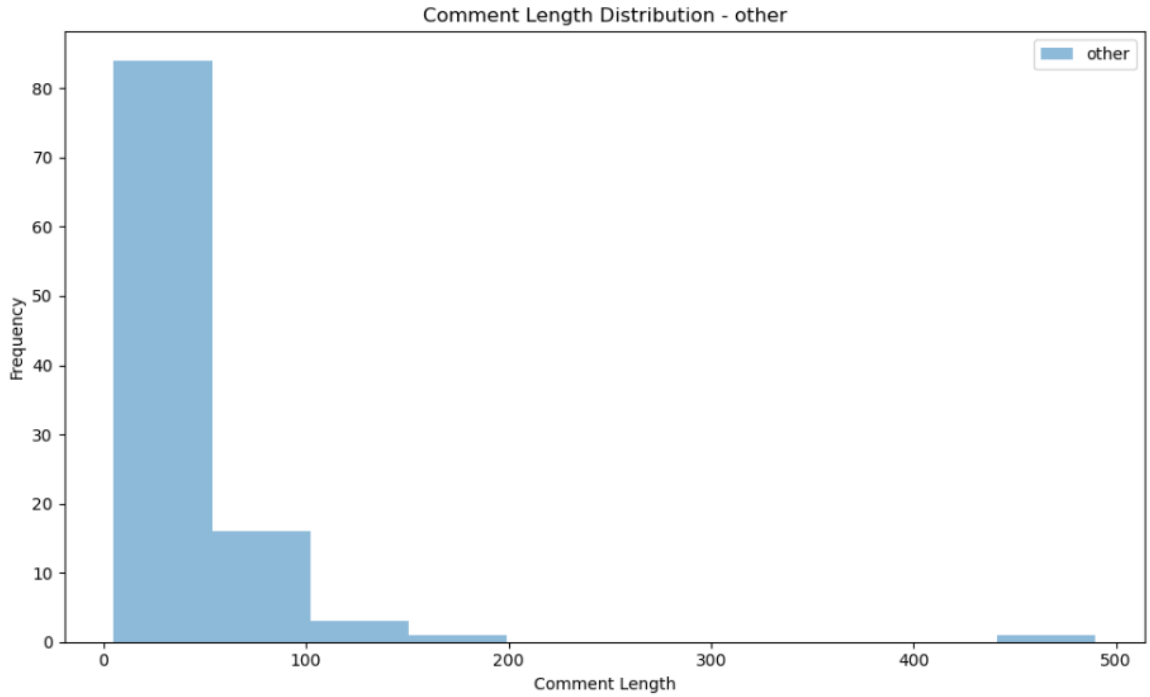
2. Apply your best classifier to each post to annotate mechanically the language of each comment and explore the annotated data. (Hint: use visualizations and extract insightful findings that would not be visible without your mechanical annotations.) A report named report.pdf should comprise these.

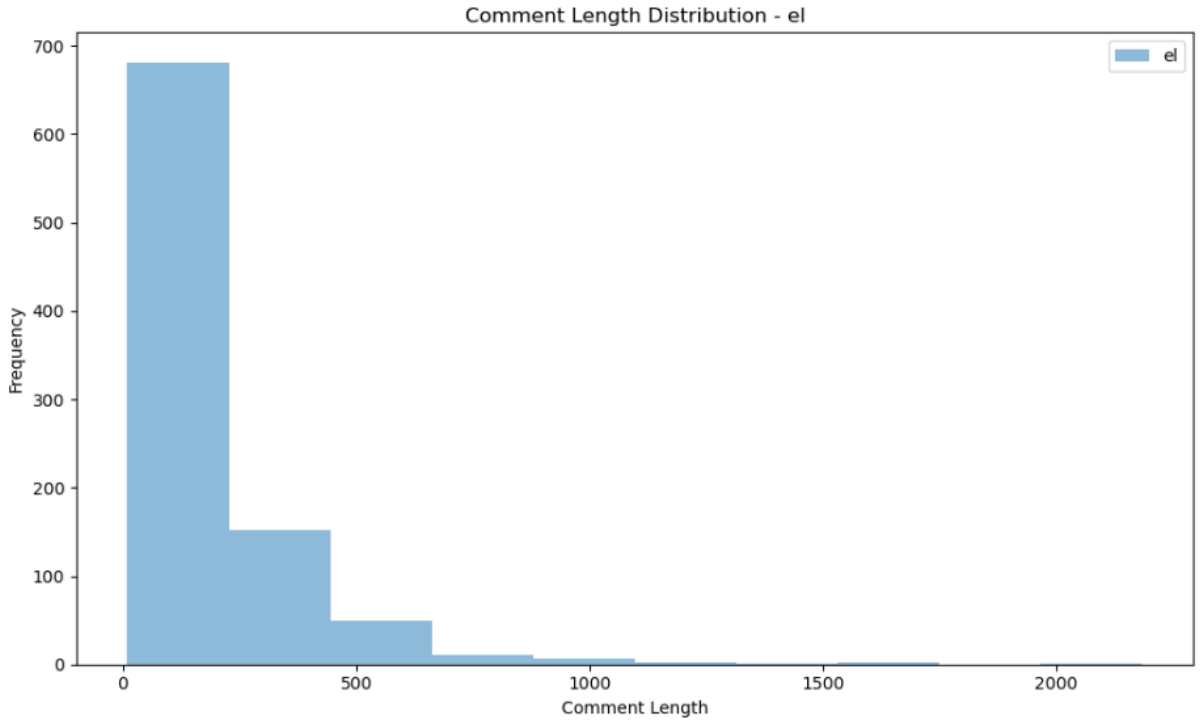


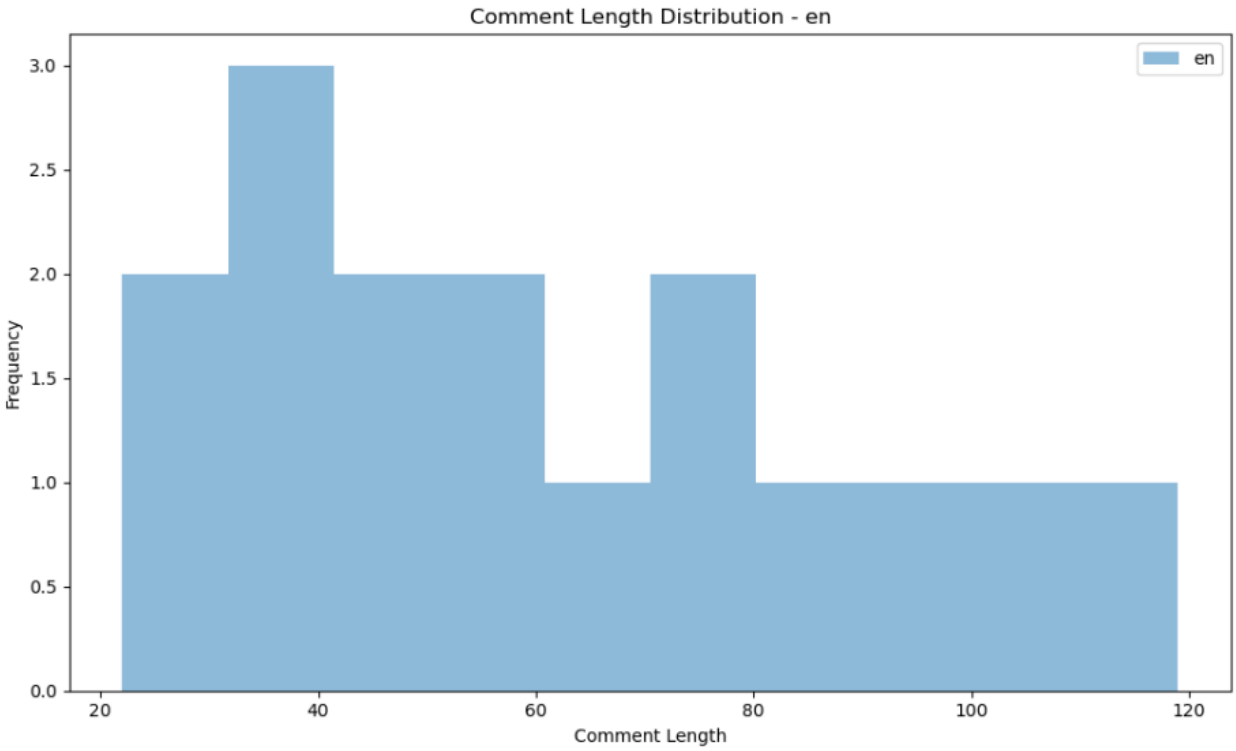
Firstly, we need to see a simple pie chart to understand the distribution of the languages in the comments. We can see that almost 90% of the comments are Greek, while only 2% are English. The first insight that we achieve is that the dialogue, in the area of YouTube that we crawled, was between Greek people, typing in their language.



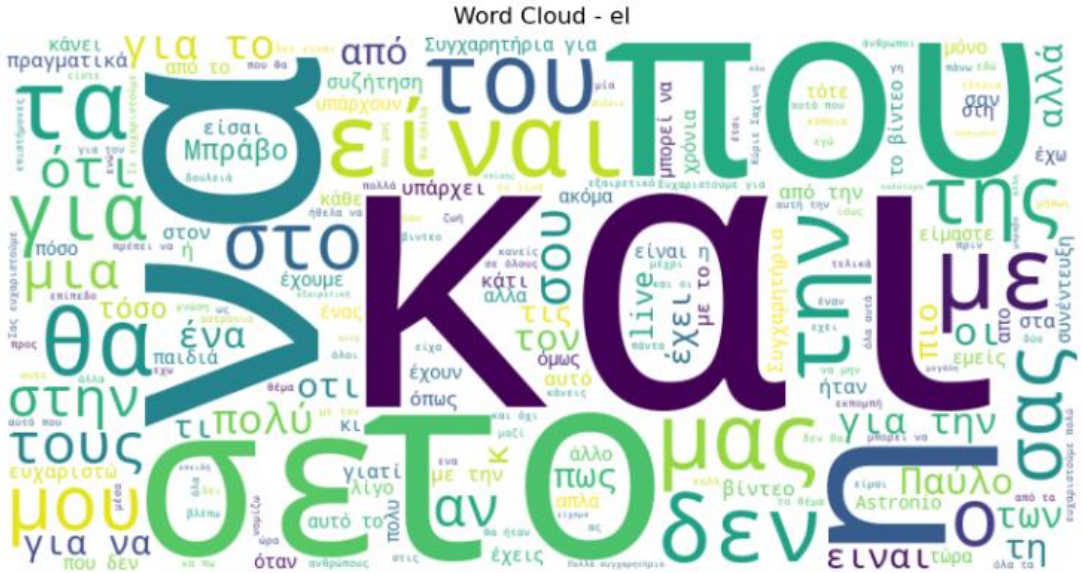
This is a simple graph to understand the language distribution in every video by also seeing the most commented videos. We can see that Astronio live has gathered the most comments in their videos including a significant percentage of different language comments. If we look closely at those videos' titles, we can see that there is a geopolitical, historical and technological aspect to their topics.







Above we can see the distribution of comment length per language. What we witness is a significantly larger size of the Greek comments, with the distribution skewed to the right, having more comments between 0 and 500 words. Other languages are distributed similarly, but with most comments being between 0 and 100 in length. We can see that English comments have a closer to uniform distribution between 20 and 120, but that could be attributed to the low sample size. The insight gained from this graph is the significantly larger comment length found in Greek comments.



Above we can see the most common words in the comments for each language. For Greek the result is expected, as small and very common words are the most usual, like και, να, που, σε, το … The most interesting insights are in the English language where we can see marble, sculptures, Sounak, belong and Parthenon. This clearly shows us that in the videos we got, there was conversation about the resent events between the Greek and English PMs around the topic of the Parthenon marbles that lie in Britain. As we said above, the most commented videos had geopolitical topics, and this could be a reason behind such a political debate in the comments.

To conclude, we can see that most of our comments were from Greek people in videos with important geopolitical topics, discussing the Parthenon marbles problem and the diplomatic problems between Greece and Britain. We can understand that there is debate or strong conversation due to the significantly larger size of the Greek comments.

IV Toxicity classification.

3. Report (in report.pdf): (a) the most toxic language, (b) the page with the more/highest rate of toxic posts, (c) the page where toxicity is uniform over time, (d) the page where toxicity increases over time.

a)

```
Greek mean toxicity rank: 1.2693156732891833
English mean toxicity rank: 1.625
Other language mean toxicity rank: 1.5523809523809524
```

The most toxic language, according to its mean toxicity rank, was Greek (Results may be influenced due to the unbalanced nature of the dataset, With Greek having 90% of the observations).

b)

	title	comment	is_toxic	toxic_comment_rate
10	Δυνατό PICK! - YouTube	1	1	1.000
2	5.000 εξωπλανήτες, αστέρας Earendel, η αναγκαι...	2	1	0.500
4	Όλγα Μαλανδράκη : Η Χανιώτισσα Επιστήμονας που...	8	3	0.375
12	Η Συνέχεια του Λαϊκού Πολιτισμού, Αντιπρόταγμ...	80	24	0.300
3	Adapting στο game = PERMA ΚΑΛΟΣ - YouTube	4	1	0.250

This is a list of the top 5 videos, with the highest toxic comment rate. The most interesting could be the 4th as it ranks this high among toxicity having a good number of comments, while others mostly have under 5 comments (effecting their toxicity rate). So, to answer the question of the most toxic video, we could confidently say that the video titled “Η Συνέχεια του Λαϊκού Πολιτισμού, Αντιπρόταγμα στην Παγκόσμια Ομογενοποίηση – YouTube” is the best fit. (A comment was considered toxic if it had a toxicity ranking higher or equal to 3)

c)

	title	comment	toxicity_rank
5	Όταν έχεις συμπαίκτη είναι πιο EASY! - YouTube	2	0.0
9	Δημ. Δαββέτας : Ο Ναπολέων, η ταινία Σκοτ, η ...	7	0.0
13	Κεντρικό δελτίο ειδήσεων 28/11/2023 | OPEN TV ...	4	0.0
18	Ταξίδια στον Χωροχρόνο (ft. Δρ Γιώργος Παππάς,...	100	0.1
0	2. Θεωρίες συνωμοσίας και επίπεδη Γη | SciTalk...	100	0.2

We considered that the videos with the less standard deviation on toxicity scores to be the ones with the most uniform toxicity. The column toxicity_rank describes that std for each video. Again, we can see the top 5 videos with the least std. The videos on the spots 4 and 5 better reassemble uniform values of toxicity, having very low std (close to 0) while having 100 observations each.

d) We couldn’t observe an overtime increasing pattern of toxicity in any of our crawled videos. What was witnessed were sudden spikes in toxicity, which after returned to their normal levels. (as seen in the example below).

Δελτίο Ειδήσεων ATTICA TV - 28/11/2023 | ATTICA TV - YouTube