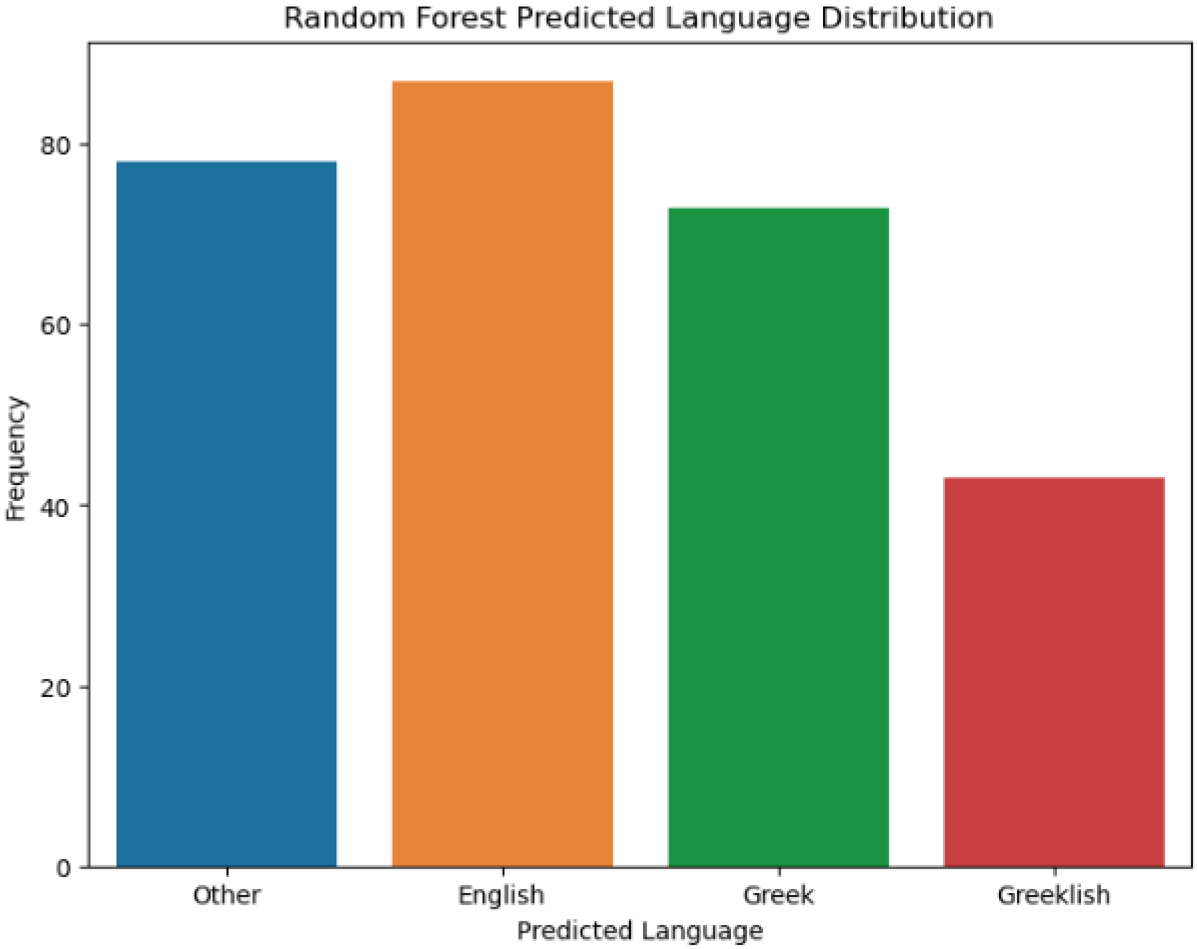
Naïve Bayes Predicted Language Distribution

In the following plot, we observe that the label 'English' has the highest frequency among the predicted labels by the Random Forest classifier. This is quite unexpected, considering that the crawled comments primarily originated from Greek and Greeklish YouTube pages.



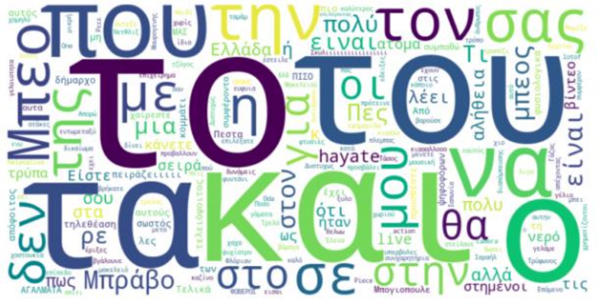
There are a couple of reasons for such an outcome, such as:

The nature of the dataset on which the classifiers were trained differs significantly from the comments to which it was applied. This suggests a potential disparity between the words used to train the classifier and those present in the comments on which we applied this model.

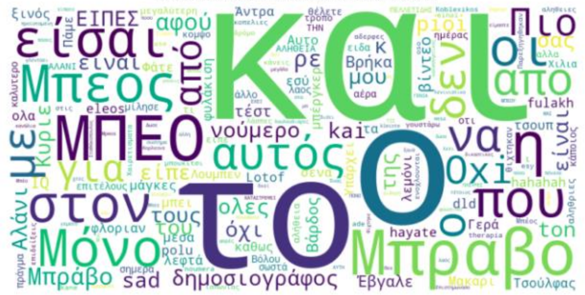
Even though the model has trained on a fairly balanced dataset, it encounters difficulty in differentiating the 'Other' and English labels from the rest of the labels. This challenge may arise because the training data labeled as 'Other' consists of a mixture of up to 10 or more languages, leading to associations between the 'Other' label and words from the other three labels: English, Greek, and Greeklish.

Word Cloud

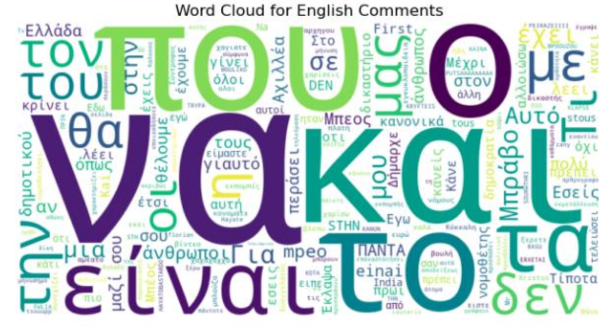
In the following series of Word Cloud screenshots, we can observe the most frequent words selected to be labeled by the respective language.



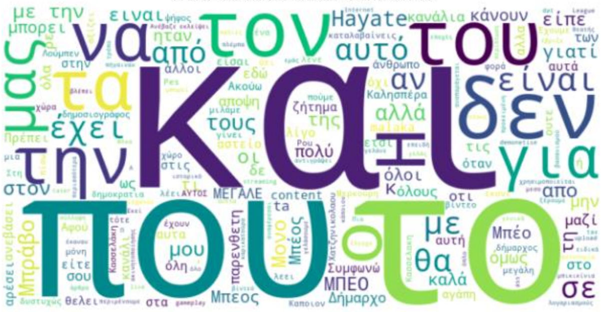
Word Cloud for Greek Comments



Word Cloud for Greeklish Comments



Word Cloud for English Comments



Word Cloud for Other Comments

It is evident that most of the labeled words belong to the Greek language, which aligns with the expected outcome given the selection of predominantly Greek and Greeklish comments. However, upon closer inspection of the words themselves, the classifier appears to have a high probability of classification errors, mislabeling many actual Greek words as English or words from other languages.

These findings complement the earlier observations that the training data associated with the 'Other' label introduces significant confusion and classification errors to our model.

It should be noted that, if we were to remove the 'Other' labeled training data, there would still be a significant overlap between the remaining classes: ['Greek', 'Greeklish', 'English']. This consideration is solely based on the observation that English-classified comments mostly contained Greek words, as the visual representation suggests. The same overlap appears in Greeklish's most frequent words, which can be considered a logical outcome, given that Greeklish is not just English letters explaining words, but rather a mixture of Greek, English language, and Greeklish typing for a given Greeklish comment.

All in all, it seems that the uniqueness of each language requires a more sophisticated approach for each language, involving a more strategic selection of training data and hyperparameter tuning if we were to upgrade this Naïve Bayes classification model.

IV Toxicity classification

(3)

(a)

After the calculation in (A2_IV – Jupyter Notebook) we came to the following results:

```
  Language: Other, Number of toxic comments: 33
  Language: Greek, Number of toxic comments: 12
 Language: English, Number of toxic comments: 25
Language: Greeklish, Number of toxic comments: 19
```

So, the most toxic comments are found on the 'Other' label indicating that the comments classified as Other are the most toxic.

(b):

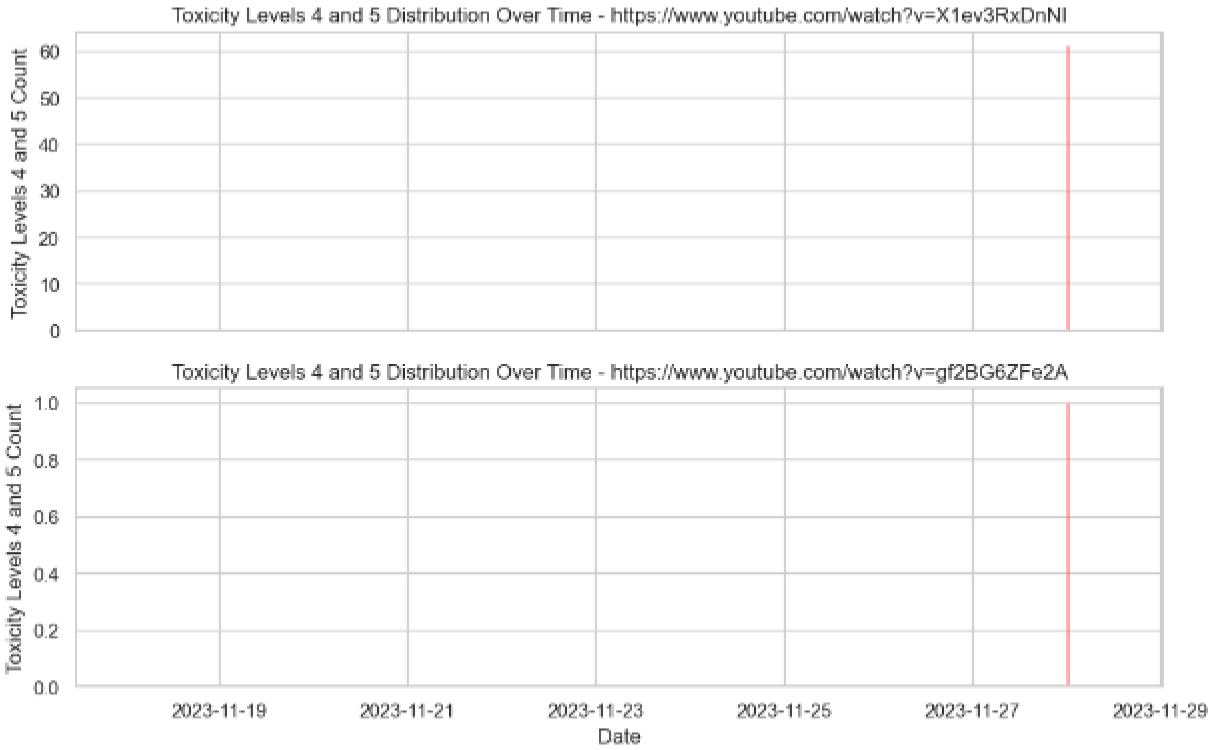
Based on our calculations the most toxic page is: https://www.youtube.com/watch?v=X1ev3RxDnNI with a toxicity rate of 35%.

With the pages: https://www.youtube.com/watch?v=vSMMBjg_CKc and https://www.youtube.com/watch?v=avB5OtMKQRc being a close second with both having a toxicity rate of 34%.

(c) ,(d):

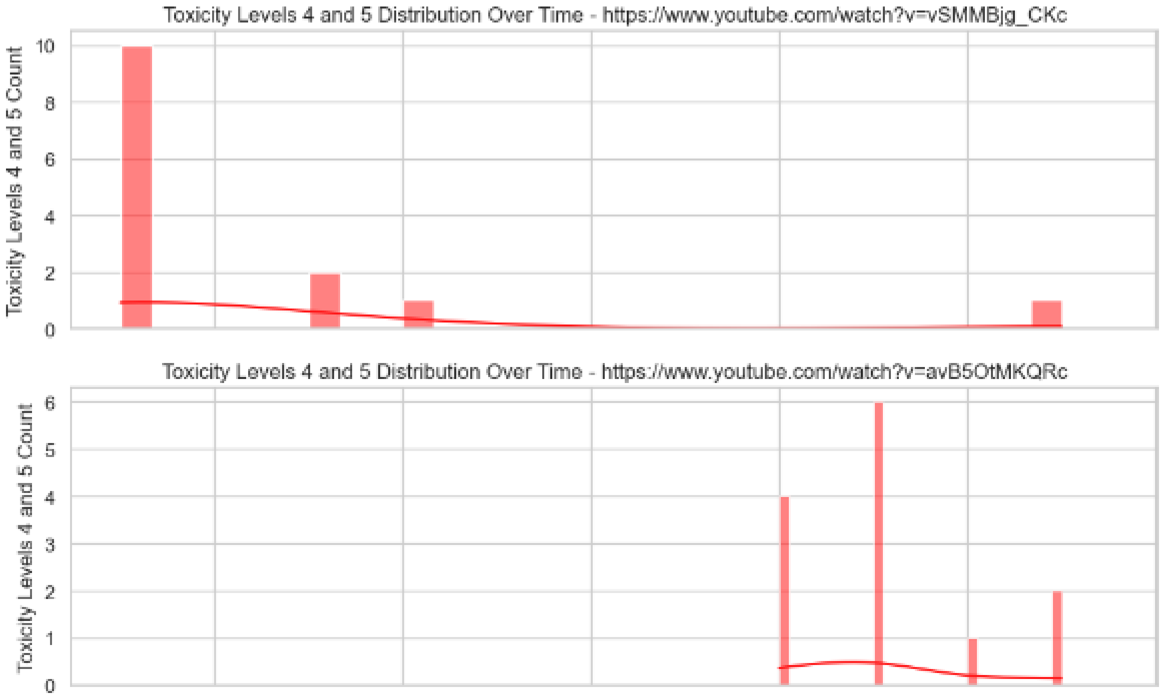
Following there will be screenshots of plots created in A2_IV Jupyter Notebook for the toxicity based on the posting date of the comments for each page.

Note-1: The distribution of the comments varies extensively in some videos and there is visible over concentration of comments in a specific date. For example, the following plots indicate that the toxic comments were closely posted on the same date:



Now, regarding the distribution of comments in first-following plot throughout the month of November 2023, we can confidently say that, on a monthly time scale, the toxicity starts from a high point in the first plot and ends at a low point, indicating a decrease in toxicity. However, this decrease is distinct and not continuous, meaning that it can drop to 0 and then increase suddenly in value.

In the second plot we can observe a sudden distribution of an increasing and decreasing number of toxic comments through out the span of a few days.



All in all, none of the provided pages display a uniform distribution of toxicity over time, but we can observe sudden changes in toxicity in the span of few days.