Supplemental Material Part 1: Crawled Data Analysis

Practical Data Science: 2nd Project

Tsirmpas Dimitris

November 29, 2023

Athens University of Economics and Business
MSc in Data Science

1 Introduction

This report outlines results and conclusions drawn from the mechanical annotation of Greek and Greeklish YouTube comment data. The full report, detailing goals, data sources, methodology and implementation can be found at `https://github.com/dimits-exe/practicaldatascience`.

Our operational data have been collected by the top results of YouTube search for two categories: **Greek songs** and **Greek Gaming videos**. These categories represent two main demographics of the Greek online community: The older - middle-aged and the younger generations.

2 Language Identification Results

Below we present graphs resulting from the language identification analysis on the crawled YouTube dataset.

- Figure 1 illustrates the distribution of languages in our YouTube dataset. Unfortunately, our classifier did not perform as expected, showing a smaller preference for Greek and Greeklish than anticipated. This discrepancy may be attributed to insufficient training data or a lack of a representative sample from the crawled posts.

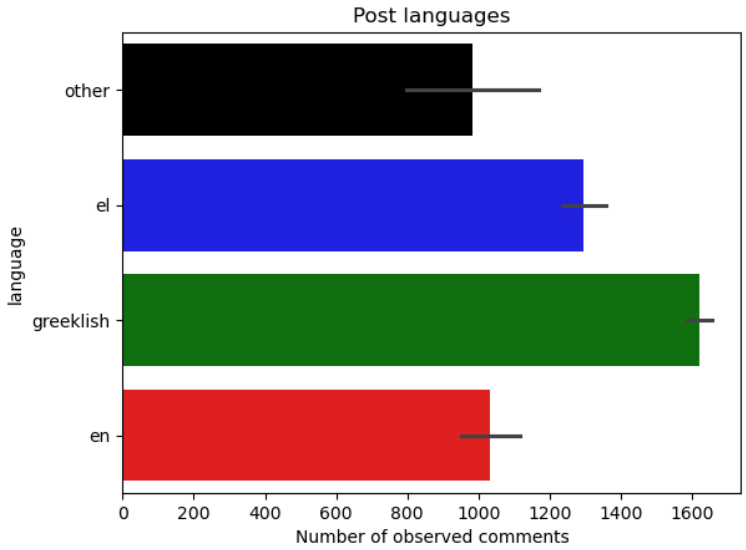

Figure 1: Language distribution in the crawled dataset.

- Figure 2 displays the number of emojis used in each comment, organized by comment language. Note that the plots are stacked, with each language on top of the previous one. To obtain the actual count for a specific language, deduct the count of emojis used in the previous language.
- Figure 3 showcases the distribution of comment lengths for each language. An outlier is observed in an English comment that reached an unusually long 1500 characters. In general, Greek posts appear shorter, while English and Greeklish dominate the longer tail of the distribution.
- Figure 4 features a time-series plot detailing the observed languages over dates. All languages exhibit a similar trend, indicating a probable absence of a statistically significant pattern. Most of the crawled videos, appearing in the YouTube search tab, are likely both relevant and popular. This explains why the majority of them are from the timeframe of 2020-2023.

3 Toxicity Classification Results

In this section we include the results of the toxicity analysis on the crawled data. The toxicity levels are defined as a scale from 1 to 5 where 1: Not Toxic, 2: Maybe Toxic, 3: Almost Toxic, 4: Almost certainly Toxic, 5: Certainly Toxic.

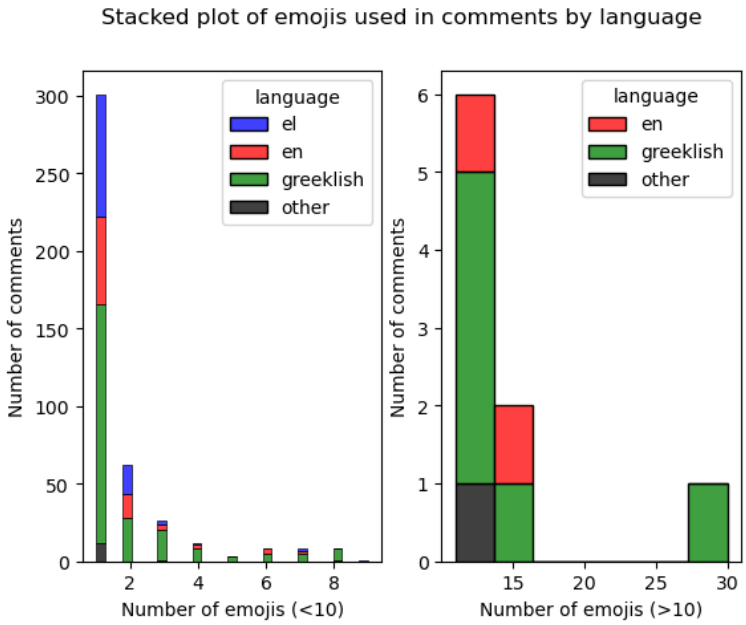

Figure 2: Stacked plot of emoji usage by language.

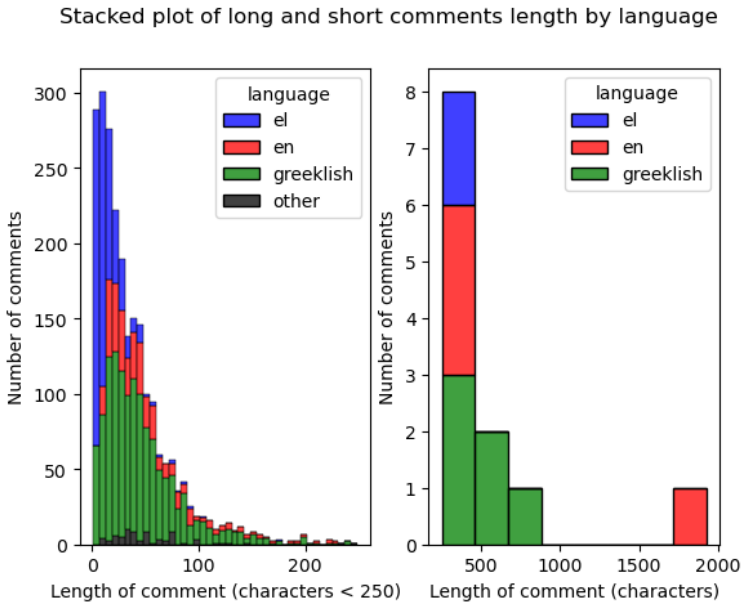

Figure 3: Length of comments by language, measured by characters.

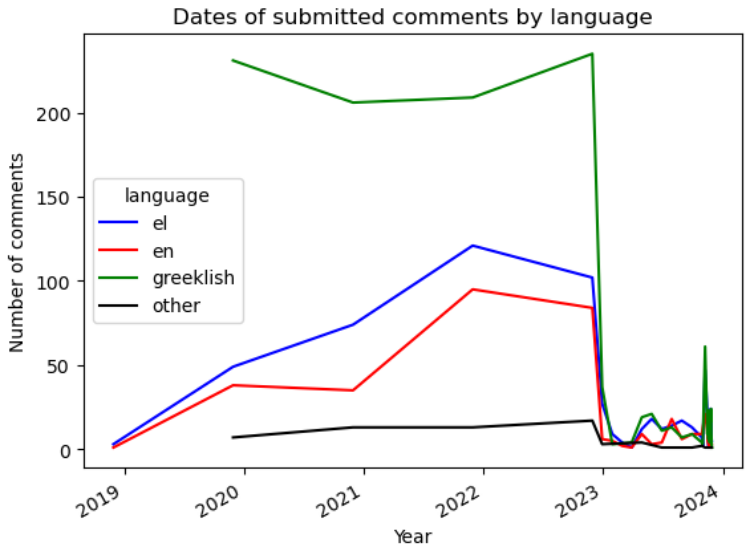

Figure 4: Number of comments by language for each day, running from 2017 to present.

Table 1: Average toxicity by language.

language	toxicity
el	1.003
en	1.118
greeklish	1.005
other	1.000

- Table 1 shows a breakdown of average comment toxicity per language. We would have expected a much bigger rating for Greeklish given the demographic and the training data.
- Table 2 shows the videos with the most toxic comments, sorted by descending toxicity. This is also a good indication that our classifier is not up to his task, as explained in the accompanying Notebook.
- Table 3 shows (a sample of) the videos where the comment toxicity remained uniform across time.
- Table 4 shows the videos where the comment toxicity increased over time, as well as the average rate of increase.

Table 2: The top 5 videos with the most toxic comments on average.

title	toxicity
Ρέμος: Αυτά όχι σε μένα, άσε τις μαγκιές	1.500
Αγαπημένα Ελληνικά Τραγούδια / Greek Music Non-Stop Mix	1.429
ΟΤΙ ΒΡΩ ΣΤΟ ORTNIT ΤΟ ΤΡΩΩ CHALLNG! (ortnite Greek)	1.286
Δεν Μιλούν Για Τον Μπρούνο (Από το "Ενκάντο: Ένας Κόσμος Μαγικός")	1.250
Ένα τραγούδι η ζωή μας - 70 αγαπημένα τραγούδια (by lias)	1.167

Table 3: Videos where comment toxicity stayed uniform over time.

title	toxicity
Greek Music Mix 2021 - Ελληνικα Τραγουδια Mix 2021 - Summer Video Greece K - Part 2	1.000
ΠΩΣ ΕΧΑΣΑ 100€ ΣΤΟ ORTNIT! *15.000 VBUCKS* (ortnite Greek)	1.000
Ένα τραγούδι η ζωή μας - 70 αγαπημένα τραγούδια (by lias)	1.000
ΚΑΘΕ KILL ΑΛΛΑΖΩ ΠΛΗΚΤΡΟΛΟΓΙΟ CHALLNG! (ortnite Greek)	1.000
Ελληνικό Έντεχνο - λαφρολαϊκό mix	1.000
ΝΙΚΗ ΜΟΝΟ ΣΤΑ ΧΙΟΝΙΑ CHALLNG ΣΤΟ ORTNIT!	1.000
ΠΡΟΚΑΛΕΣΑ STRAMSNIPR ΣΕ 1V1?! ΔΕΙΤΕ ΤΙ ΕΓΙΝΕ..	1.000
Αναστασία & Josephine - Είσαι Μια Θεά — Mad Video Music Awards 2023 από τη ΔΕΗ	1.000
ΞΕΚΛΕΙΔΩΣΑ ΟΛΟ ΤΟ SASON 6 BATTL PASS! (ortnite Greek)	1.000
ΗΡΘΕ Η ΝΕΑ ΤΡΕΛΗ ΣΕΖΟΝ ΣΤΟ ORTNIT!	1.000
ΝΙΚΗ ΜΟΝΟ ΣΤΑ ΧΙΟΝΙΑ CHALLNG! (ortnite Greek)	1.000
Νίκησα Χωρίς Να Προσγειωθώ Στο OG ortnite!	1.000
Έντεχνα & Λαϊκά - Τα Τραγούδια Που Αγαπάμε [1]	1.000
ncanto αλά ελληνικά — NeverLander	1.000
Δεν Μιλούν Για Τον Μπρούνο (Από το "Ενκάντο: Ένας Κόσμος Μαγικός")	1.000
Ψίθυροι θάρρους. — Whispers of Courage in Greek — @GreekairyTales	1.000
ΛΑΪΚΑ MIX — KONSTANTINOS SOT	1.000
ΝΙΚΗ ΜΕ ΤΟΝ ΤΡΟΧΟ ΤΗΣ ΤΥΧΗΣ CHALLNG! (ortnite Greek)	1.000
ΝΙΚΗ ΜΕ ΤΗΝ *ΣΠΑΝΙΑ* POZ GHOUL TROOPR (ortnite Greek)	1.000
ΞΕΚΛΕΙΔΩΣΑ ΟΛΟ ΤΟ SASON 9 BATTL PASS! (ortnite Battle Royale)	1.000
Συμβουλές και Κόλπα για να γινεις Καλύτερος παιχτης! (ortnite Greek)	1.000
10χρονος Παίρνει Νίκη Στο OG ORTNIT	1.000
ΑΝ ΓΕΛΑΣΕΙΣ ΧΑΝΕΙΣ 500 VBUCKS! (ortnite unny Moments)	1.000
ΤΡΑΓΟΥΔΙΑ ΠΟΥ ΑΓΑΠΗΣΑΜΕ (ΣΥΛΛΟΓΗ ΜΕ ΕΛΛΗΝΙΚΑ ΤΡΑΓΟΥΔΙΑ)	1.000
ΤΟ OG ORTNIT ΕΠΕΣΤΡΕΨΕ!! ft Alekkun MrAntouon	1.000
Ελληνικά disco, 80s & 90s (Non-stop Party Mix) [Part 1]	1.000
Αν θα μπορούσα τον κόσμο να άλλαζα - 30 έντεχνα τραγούδια που αγαπάμε (by Linda)	1.000
ΝΙΚΗΣΑΜΕ ΣΤΟ DADPOOL CHALLNG! (ortnite Greek)	1.000
0 ελληνικά τραγούδια από τα 60's (by lias)	1.000
Josephine - Μοίρα - Official Music Video	1.000

Table 4: Videos where comment toxicity stayed increased over time.The toxicity_diff represents the average difference between comment toxicitywith lag 1 across each date.

title	toxicity_diff
Greek Music Mix 2021 - Ελληνικα Τραγουδια Mix 2021 - Summer Video Greece K - Part 2	3.000
Αγαπημένα Ελληνικά Τραγούδια / Greek Music Non-Stop Mix	3.000
Ρέμος: Αυτά όχι σε μένα, άσε τις μαγκιές	3.000
ΣΚΟΤΩΣΑ ΤΟΝ MONGRAAL M 2OBOMB !	3.000
ΑΝ ΓΕΛΑΣΕΙΣ ΧΑΝΕΙΣ 500 VBUCKS! (ortnite unny Moments)	3.000
Δοκίμασα 30 PS στο UNRAL RANK...	3.000
Ελαφρολαϊκά παλιά - 120 μεγάλες επιτυχίες (by Linda's Music Dream)	3.000
Ένα τραγούδι η ζωή μας - 70 αγαπημένα τραγούδια (by lias)	2.500
Δεν Μιλούν Για Τον Μπρούνο (Από το "Ενκάντο: Ένας Κόσμος Μαγικός")	2.333
ΝΙΚΗ ΜΟΝΟ ΜΕ ΜΥΘΙΚΑ ΟΠΛΑ CHALLNG! (ortnite Greek)	2.000
ΕΒΓΑΛΑΝ ΤΟ *BUILDING* ΣΤΗΝ ΝΕΑ SASON ΤΟΥ ORTNIT! (ortnite Greek)	2.000
Νίκη ΜΟΝΟ με Πιστόλι Challenge (ortnite OG)	2.000
ΠΩΣ ΕΧΑΣΑ 100€ ΣΤΟ ORTNIT! *15.000 VBUCKS* (ortnite Greek)	2.000
ΟΤΙ ΒΡΩ ΣΤΟ ORTNIT ΤΟ ΤΡΩΩ CHALLNG! (ortnite Greek)	2.000
GRK 2K23 SUMMR MIX — VOL. I — by NIKKOS DINNO — ΗΡΘΕ ΚΑΛΟΚΑΙΡΙ —	2.000