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.

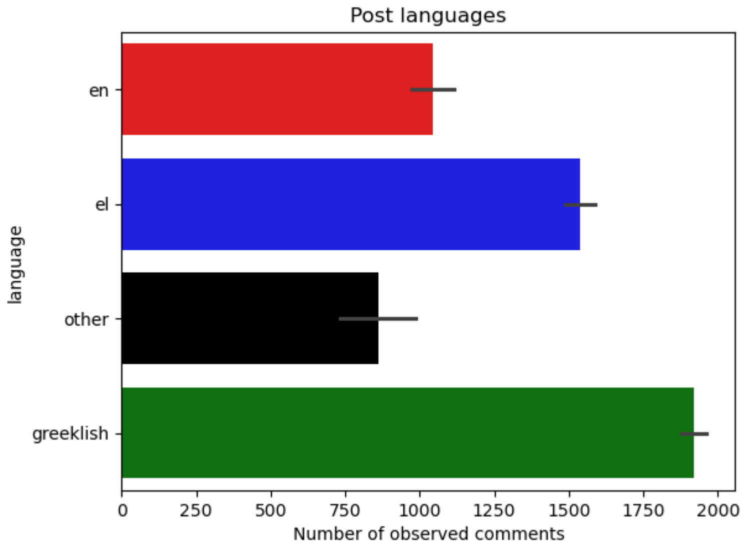

Figure 1: Language distribution in the crawled dataset.

- 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.

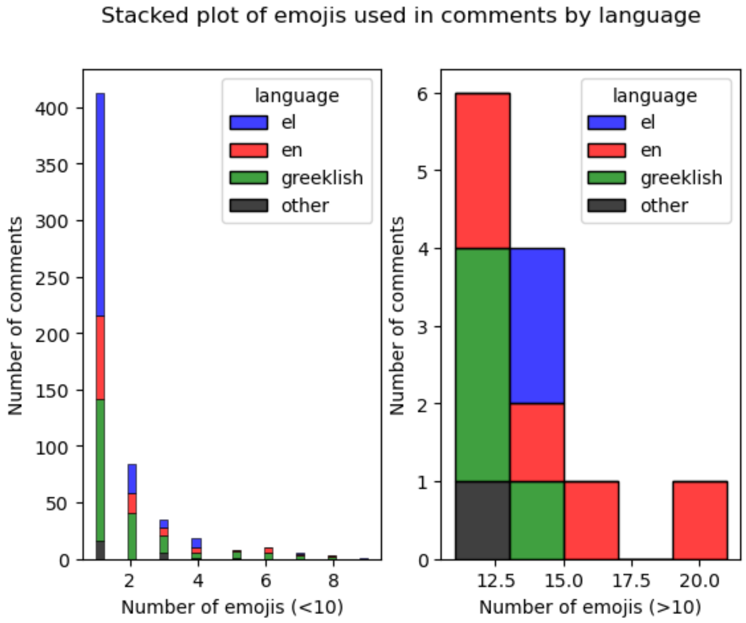

Figure 2: Stacked plot of emoji usage by language.

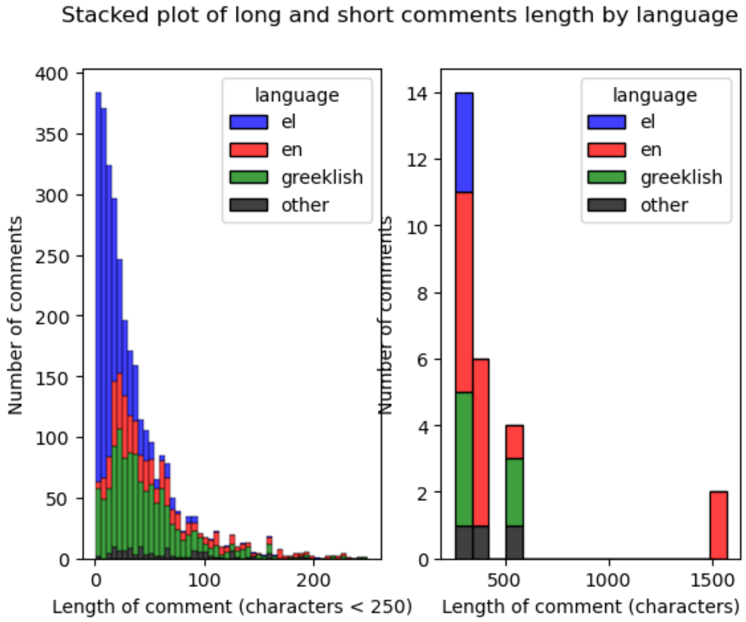

Figure 3: Length of comments by language, measured by characters.

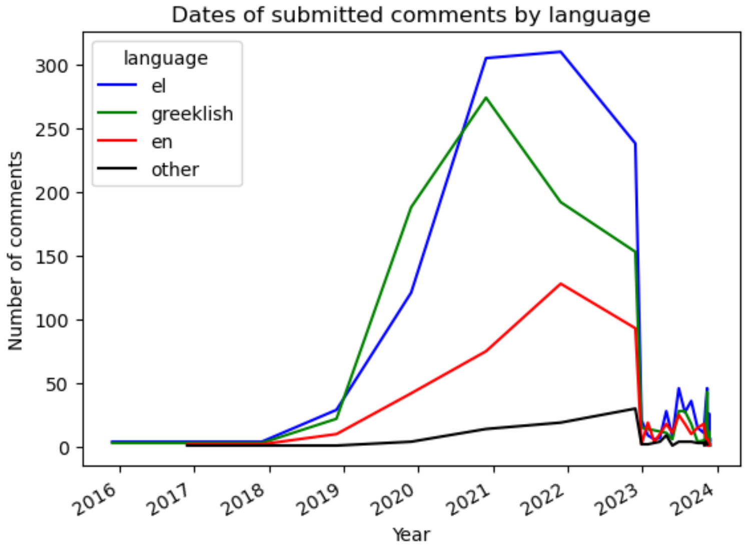

Figure 4: Number of comments by language for each day, running from 2017 to present.

Table 1: Average toxicity by language.

language	toxicity
el	1.000
en	1.088
greeklish	1.011
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
Ρομαντικά Ελαφρά Τραγούδια — Non Stop Mix	1.333
ΞΕΚΛΕΙΔΩΣΑ ΟΛΟ ΤΟ BATTL PASS ΤΗΣ SASON 2! (ortnite Greek)	1.214
Ελαφρολαϊκά παλιά - 120 μεγάλες επιτυχίες (by Linda's Music Dream)	1.136
Nina Mazani - Άγχος (Από το "Ενκάντο: Ένας Κόσμος Μαγικός")	1.133
Ένα τραγούδι η ζωή μας - 70 αγαπημένα τραγούδια (by lias)	1.122

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

title	toxicity
ΤΑ ΣΠΑΜΕ ΕΛΛΗΝΙΚΑ — KONSTANTINOS SOT	1.000
5 Ωρες Non Stop special!! Αποκλειστικά για μερακλήδες! Στην υγειά μας κ Όλοι οι καλοί χωράνε.	1.000
Ελληνικό Έντεχνο - λαφρολαϊκό mix	1.000
ΠΡΟΚΑΛΕΣΑ STRAMSNIPR ΣΕ 1V1?! ΔΕΙΤΕ ΤΙ ΕΓΙΝΕ..	1.000
ΤΡΟΛΛΑΡΩ ΚΙΝΕΖΟΥΣ ΣΤΟ ORTNIT! (ortnite Greek)	1.000
ΞΕΚΛΕΙΔΩΣΑ ΟΛΟ ΤΟ SASON 6 BATTL PASS! (ortnite Greek)	1.000
ΤΡΑΓΟΥΔΙΑ ΠΟΥ ΑΓΑΠΗΣΑΜΕ (ΕΠΙΛΟΓΕΣ)	1.000
KAPS TO MAGAZI 2K23 [Bouzoukia Live Mix II] by NIKKOS DINNO — Ελληνικά Μπουζούκια	1.000
Έντεχνα & Λαϊκά - Τα Τραγούδια Που Αγαπάμε [1]	1.000
Greek hits 90's mix / ΕΛΛΗΝΙΚΕΣ ΕΠΙΤΥΧΙΕΣ NON STOP	1.000
Ρομαντικά Ελαφρά Τραγούδια — Non Stop Mix	1.000
ΔΕΙΧΝΩ ΓΙΑ ΠΡΩΤΗ ΦΟΡΑ ΤΑ SKINS ΜΟΥ !!!	1.000
ΝΙΚΗΣΑΜΕ ΣΤΟ DADPOOL CHALLNG! (ortnite Greek)	1.000
ΝΙΚΗ ΜΕ ΤΗΝ *ΣΠΑΝΙΑ* ΡΟΖ GHOUL TROOPR (ortnite Greek)	1.000
ΕΠΑΙΞΑ ΤΟ ΑΛΗΘΙΝΟ ΚΙΝΕΖΙΚΟ ORTNIT!	1.000
ΞΕΚΛΕΙΔΩΣΑ ΟΛΟ ΤΟ SASON 9 BATTL PASS! (ortnite Battle Royale)	1.000
100 Λαϊκά Χορευτικά - 100 Laika Horeftika — Non Stop Mix	1.000
ΑΝ ΓΕΛΑΣΕΙΣ ΧΑΝΕΙΣ 500 VBUCKS! (ortnite unny Moments)	1.000
ΤΕΛΙΚΟΣ ORTNIT WORLD CUP - ΜΕΡΑ 3 SOLOS (Επίσημο Ελληνικό Show)	1.000
ΞΕΚΛΕΙΔΩΣΑ ΟΛΟ ΤΟ BATTL PASS ΤΗΣ SASON 2! (ortnite Greek)	1.000
Μια Αγγλικη λεξη = Αφρος!! {ortnite Greek}	1.000
Ελαφρολαϊκά παλιά - 120 μεγάλες επιτυχίες (by Linda's Music Dream)	1.000
ΣΚΟΤΩΣΑ ΤΟΝ MONGRAAL M 2OBOMB !	1.000
Πασχάλης Τερζής, Βασίλης Καρράς, Νότης Σφακιανάκης - 36 αγαπημένα τραγούδια — No.1 (by Linda)	1.000
ΠΑΙΖΟΥΜΕ ONLY UP ΣΤΟ ORTNIT!	1.000
Γιώργος Λιβάνης & Dirty Harry - Αν Στο Χέρι Σου Είναι - Official Music Video	1.000
0 ελληνικά τραγούδια από τα 60's (by lias)	1.000
OG SASON 6 UPDAT ΣΤΟ ORTNIT ΠΑΜΕ ΝΑ ΤΟ ΤΣΕΚΑΡΟΥΜΕ !!!	1.000
Όμορφα ελληνικά τραγούδια	1.000
Συμβουλές και Κόλπα για να γινεις Καλύτερος παιχτης! (ortnite Greek)	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
ΜΕ SCAMMAR ΚΟΡΙΤΣΙ(SCAMMR GTS SCAMMD){GRK}	3.000
Πέρασα 2 ΩΡΕΣ στο OG ortnite	3.000
ΣΚΟΤΩΣΑ ΤΟΝ MONGRAAL M 2OBOMB !	3.000
ΞΕΚΛΕΙΔΩΣΑ ΟΛΟ ΤΟ BATTL PASS ΤΗΣ SASON 2! (ortnite Greek)	3.000
ΑΝ ΓΕΛΑΣΕΙΣ ΧΑΝΕΙΣ 500 VBUCKS! (ortnite unny Moments)	3.000
ΗΡΘΕ ΤΟ HALLOWN ΣΤΟ ORTNIT ft W1ndz	3.000
Ελαφρολαϊκά παλιά - 120 μεγάλες επιτυχίες (by Linda's Music Dream)	3.000
5 Ωρες Non Stop special!! Αποκλειστικά για μερακλήδες! Στην υγειά μας κ Όλοι οι καλοί χωράνε.	2.600
Ένα τραγούδι η ζωή μας - 70 αγαπημένα τραγούδια (by lias)	2.500
Nina Mazani - Άγχος (Από το "Ενκάντο: Ένας Κόσμος Μαγικός")	2.000
Κωνσταντίνος Αργυρός - Ελπίδα - Official Music Video - Konstantinos Argiros "lpida"	2.000
ΠΑΛΙΑ ΛΑΙΚΑ ΓΙΑ ΟΛΑ ΤΑ ΓΟΥΣΤΑ!!!!!MIX - ΕΠΙΛΕΓΜΕΝΑ ΛΑΙΚΑ ΤΡΑΓΟΥΔΙΑ	2.000
Ρομαντικά Ελαφρά Τραγούδια — Non Stop Mix	2.000
TRYHARDIANS O TH GALAXY ΣΤΟ ΝΕΟ ARNA MOD! ft. PanosDentGames	2.000
1v1 ΜΕ ΤΟΝ ΠΙΟ *TOXIC OG* ΠΑΙΧΤΗ ΣΤΟΝ ΚΟΣΜΟ	2.000