Supplemental Material

Practical Data Science: 2nd Project

Tsirmpas Dimitris

November 26, 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

The main questions about our dataset are:

- How often is Greeklish spoken in circles of young and old people?
- How much do Greek and English users use emojis in comments?
- Do comments in Greeklish use emojis more often than comments in Greek?
- Is there any trend towards using Greeklish in the recent years?

Below, we present supporting graphs.

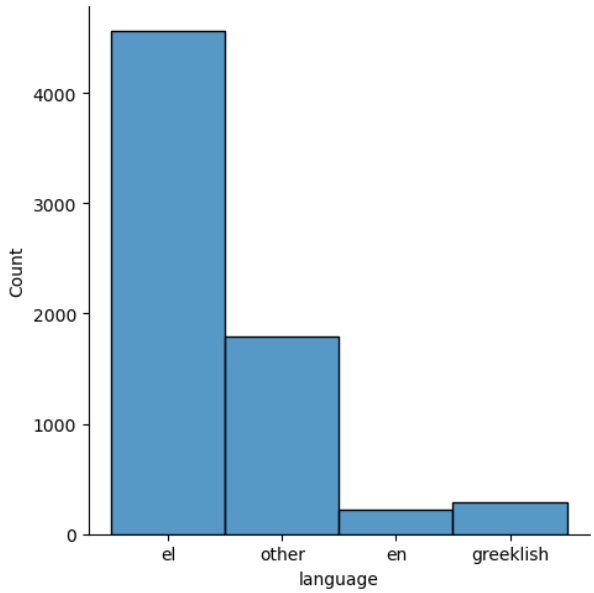

Figure 1: Language distribution in the crawled dataset.

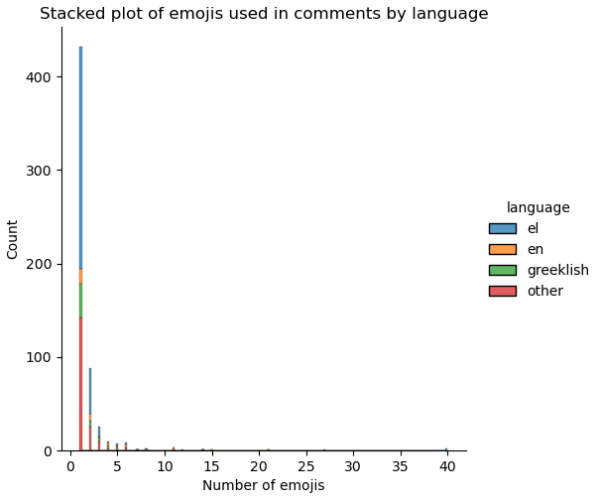

Figure 2: Stacked plot of emoji usage by language.

3 Toxicity Classification Results

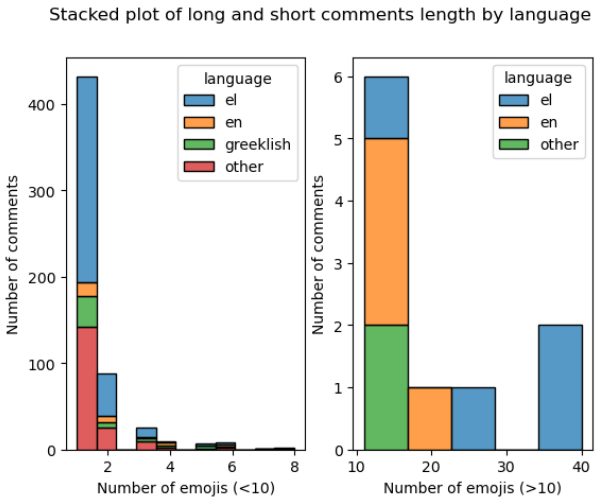

Figure 3: Length of comments by language, measured by characters.

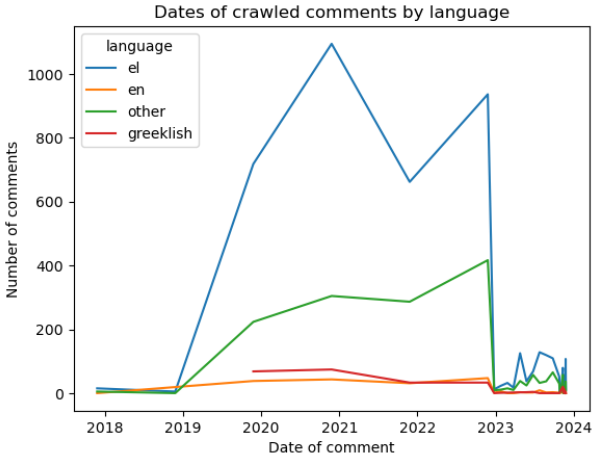

Figure 4: Number of comments by language for each day, running from 2017 to present.