Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA

PRIN 2022 Project EPICA

Deliverable 3.1
Technical report describing annotated datasets

Pietro Baroni[1], Stefano Bistarelli[2], Bettina Fazzinga[3], Giulio Fellin[1], Sergio Flesca[4], Filippo Furfaro[4], Massimiliano Giacomin[1], Francesco Parisi[4], Carlo Proietti[5], Irene Russo[5], Francesco Santini[2], Carlo Taticchi[2], and Paola Vernillo[5]

[1]DII - Universitá di Brescia

[2]DMI - Universitá di Perugia

[3]DICES - Universitá della Calabria

[4]DIMES - Universitá della Calabria

[5]ILC - Consiglio Nazionale delle Ricerche

Abstract

This technical document outlines the criteria used to build a dataset from public campaigns that promote increased fruit and vegetable consumption in multiple countries. Section 1 summarizes how relevant campaigns were selected. Section2 details data collection (web crawling) and noise-reduction cleaning procedures.

1 Campaigns' Selection Criteria

We built a specialized bilingual corpus focused on campaigns promoting greener diets, with an emphasis on increasing fruit and vegetable consumption. The corpus includes content in both English and Italian. For English-language materials, we gathered campaigns originating from countries such as the United

States, United Kingdom, Ireland, Canada, Australia, and New Zealand. For Italian-language materials, we exclusively collected campaigns designed for audiences within Italy. The data collection process spanned two months, specifically from October to December 2024.

The development of this dataset followed a two-tiered methodological approach. Initially, we conducted a manual search on Google using the query “fruit & vegetables campaign + country,” analyzing results up to the first 15 pages. This was complemented by the use of ChatGPT, employing prompts such as “Provide campaigns on fruit and vegetable consumption in + country” to identify additional resources.

To ensure the representativeness and relevance of the corpus, campaigns retrieved online were not indiscriminately included but were subjected to a rigorous selection process. Specifically, only campaigns explicitly focused on the promotion of fruit and vegetable consumption were incorporated, while broader health initiatives (e.g., those addressing salt reduction or increased physical activity) were excluded. As a result, the final dataset comprises a total of 45 campaigns, of which 41 are in English and only 4 are in Italian. The corpus is structured as follows:

- **United States**: 10 campaigns;
- **United Kingdom**: 11 campaigns;
- **Australia**: 11 campaigns;
- **Ireland**: 3 campaigns;
- **Canada**: 3 campaigns;
- **New Zealand**: 3 campaigns;
- **Italy**: 4 campaigns.

Furthermore, each campaign was classified within an Excel database according to selected key categories:

- **language**: system of communication used for the promoting campaign (i.e., English vs. Italian);
- **campaign name**: name that a particular social or marketing campaign is known by (e.g., Vegout);

- **country**: recipient state or nation of the campaign (e.g., United Kingdom);

- **level of promotion**: a) the campaign is connected to or administered by the government (governmental campaigns) vs. the campaign is independent of government and connected to or administered by a charity, association, etc. (non-governmental campaigns); b) the campaign is intended to be promoted throughout the territory (central; i.e., UK) vs. the campaign is directed to a particular place, region, or area of a wider territory (local; i.e., Yorkshire);

- **level of the organization**: the fostering institution, agency, or charity has its legal headquarters and may operate a) in one of the areas that a country is divided into (i.e., regional), b) nationwide (national), c) throughout Europe (i.e., European), or d) across multiple independent states (i.e., international);

- **name of promoting organization**: the name that a particular institution, agency, or charity is known by (e.g., Foundation for Fresh Produce or FFP);

- **type of promoting organization**: the organization is connected to or administered by the government (i.e., governmental) vs. the organization operates outside government control (i.e., non-governmental);

- **year of launch**: the time frame for the campaign launch and dissemination (e.g., from 2021 to present);

- **target audience**: the promotional campaign is built for targeting a universal audience regardless of race, gender, age, socio-economic status, or educational level (i.e., universal; adults, adolescents, and children) vs. the promotional campaign is built for targeting a particular group of people or community (i.e., specific; young adolescents and children);

- **type of communication resources**: the campaign's official website provides the targeted audience with background documents and informational resources (e.g., direct persuasive communication) vs. the background documents and informational resources are destined for third parties other than the persons accessing the campaign's official website (i.e., mediated persuasive communication; recipes and infographics for teachers and parents to be used in schools or at home); the resources are made available on the campaign official's website (i.e., general informational materials) vs. in the absence of an official website, campaign informational resources are made available within a dedicated section of the fostering organization's website (i.e., campaign-specific documents).

2 Web Crawling and Data Cleaning

The textual content of the websites related to the selected campaigns was extracted via crawling. This script performing crawling is a depth-1 web harvester for public-interest campaign sites: it launches Chrome via Selenium, loads the homepage, collects all deduplicated HTTPS links, then visits each once to grab the rendered HTML, strip boilerplate (script/style), extract visible text, split it into sentences (using `NLTK`), and save the extracted content in a CSV of sentence-level records with campaign_id, sentence, link, name_campaign. It relies on `BeautifulSoup` for parsing and pandas for I/O. The crawl is intentionally shallow (homepage → first-level links), with no recursion, no robots.txt checks, and no cookie-banner handling. It produces a reproducible, sentence-granular corpus from a campaign site for downstream analysis and annotation.

The crawling process successfully extracted textual content from 38 of the 45 previously selected campaigns, as some websites blocked the automated procedure. After cleaning the data (removing duplicates and discarding very short sentences of fewer than 5 words), the resulting dataset contained 29,864 sentences. This dataset still includes some irrelevant or noisy information for the EPICA project case study (e.g., addresses, name lists, technical documentation). Therefore, it must be further cleaned to be processed by the pipeline described in Deliverable 3.3.

To identify sentences with potentially argumentative content that are suitable for evaluation in terms of argumentative entailment each sentence was analyzed with a model trained on domain-specific data (`bert-base_claimbuster`). The model assigns a probability score to each sentence, indicating whether it can be considered a claim. It is important to note that this probability corresponds to the model's softmax score.

`bert-base_claimbuster` is a BERT-based model fine-tuned on the ClaimBuster dataset, which consists of 23,533 fact-check worthy statements extracted from U.S. presidential election debates [1]. The model is designed for claim detection and achieves an accuracy of 82%.

References

[1] Fatma Arslan, Naeemul Hassan, Chengkai Li, and Mark Tremayne. A benchmark dataset of check-worthy factual claims. In *International Conference on Web and Social Media*, 2020.