MISSION 4
EDUCATION
RESEARCH

PRIN 2022 PNRR

Normative and Digital Solutions to Counter Threats during National Election Campaigns (RightNets)

Project Code: P2022MCYCK | CUP: D53D23022340001

University of Macerata

PI - Prof. Giovanni Di Cosimo

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

Normative and Digital Solutions to Counter Threats during National Election Campaigns (RightNets)

Deliverable 3.1 – Plan for the data collection process

Project Title	Normative and Digital Solutions to Counter Threats during National Election Campaigns
Project Acronym	RightNets
Call Identifier	PRIN 2022 PNRR - Decreto Direttoriale n. 1409 del 14-9-2022
Grant. no	P2022MCYCK
CUP	D53D23022340001
Start of Project	30 November 2023
Duration	24 months
Work Package	3
Abstract	The Data Collection Plan for RightNets aims to ensure that social media data will be systematically gathered, to study the impact of digital campaigning on democratic processes. The plan focuses on data collection from social networks, specifically Instagram and X (formerly Twitter), during the 2022 Italian Political elections and the 2024 European elections to identify patterns in political engagement, propaganda, and foreign interference. As such, this document outlines the use of two primary data sources — Instagram and X — due to their significant roles in political communication and campaigning. Two scrapers will be implemented to gather public posts containing election-related key words. Metadata, such as post IDs, account names, and likes, will be systematically collected, ensuring compliance with data source policies. Data management practices are structured to adhere to ethical standards, focusing on public content and ensuring privacy. The collected data will be processed in compliance with data protection regulations and made available on GitHub, adhering to FAIR principles.
Authors	Paolo Sernani, Emanuele Frontoni
PI	Giovanni Di Cosimo

Il Progetto *Normative and Digital Solutions to Counter Threats during National Election Campaigns – RightNets* è finanziato dall'Unione europea - Next Generation EU, nell'ambito del bando PRIN 2022 PNRR Missione 4, Componente 2, Investimento 1.1 (P2022MCYCK; CUP D53D23022340001).
I punti di vista e le opinioni espresse sono solo quelli degli Autori e non riflettono necessariamente quelli dell'Unione europea o della Commissione europea. Né l'Unione europea né la Commissione europea possono essere ritenute responsabili per essi.

Data Collection Plan

Paolo Sernani*, Emanuele Frontoni+
* Department of Law, University of Macerata, Italy, paolo.sernani@unimc.it
+ Department of Political Sciences, Communication and International Relations, University of Macerata, Italy, emanuele.frontoni@unimc.it

In WP3 of the RightNets project, the data collection purpose is to gather data from social networks to analyze digital campaigning and its impact on democratic processes. This data will enable the identification and examination of patterns related to targeted propaganda, ad spending, political engagement, and potential foreign interference. By collecting data on social media interactions, the project aims to create models that assess compliance with campaign financing laws, enhance transparency in political advertising, and explore candidates' accountability to their constituencies. As such, in this report we outline the methods and processes for gathering, managing, and analyzing data to achieve specific research objectives. It ensures that the collected data is relevant, reliable, and aligned with the study's goals. This plan is structured as follows. After this brief introduction, that defines the purpose and scope of data collection, we give an overview of dataset already published about the 2022 Italian political elections. Then, we will list the selected data sources for our collection, detailing where and how data will be obtained, i.e., the collection method: in this section, we describe the web scraping tools implemented for the project, using public APIs. Data management, in compliance with data source policies will be described next, with a timeline, specifying the stages of data collection and key milestones. The plan aims to ensure that data collection is systematic, compliant to regulations, and scientifically valid.

Existing data from social media interactions about the 2022 Italian political election

Some datasets about social media interactions regarding the 2022 Italian political election are available. To this end, ITA-ELECTION-2022[1] is a public, open access dataset of social media activity related to the 2022 Italian election. The dataset includes millions of posts from Facebook, Instagram, and Twitter, along with metadata from TikTok and YouTube. Data were collected between July and October 2022 using public APIs and keyword searches, including posts from CrowdTangle[2], a Meta tool. The dataset also features social media accounts of Italian politicians and was built using a snowball sampling approach starting with terms like "elezioni2022." Key statistics include 19,087,594 tweets, 1,142,812 Facebook posts, 68,078 Instagram posts, and metadata for 22,754 YouTube and 1,903 TikTok videos. The dataset is publicly available on GitHub[3]. Although no experiments were conducted, the dataset aims to enhance understanding of social media's impact on democratic processes. Trastulli and Mastroianni[4] compiled a dataset of 14 party manifestos from the 2022 Italian election, representing 97.7% of the votes. Parties included range from "Alleanza Verdi e Sinistra" to "Unione Popolare." Non-searchable PDFs were converted to machine-readable text using Optical Character Recognition (OCR). The text was tokenized, cleaned, and transformed into a document-feature matrix (DFM) for analysis using the R language. The study applied methods like relative frequency, collocation, and Key Word In Context (KWIC) analysis to examine socio-economic issues, administrative reforms, and key themes like environmental concerns. The findings suggest that party manifestos balanced various topics, emphasizing traditional ideological stances, with less focus on

[1] Pierri, F., Liu, G., Ceri, S.: ITA-ELECTION-2022: A multi-platform dataset of social media conversations around the 2022 Italian general election. In: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23), pp. 5386–5390. ACM, New York (2023). https://doi.org/10.1145/3583780.3615121

[2] https://www.crowdtangle.com/

[3] https://github.com/frapierri/ita-election-2022

[4] Trastulli, F., Mastroianni, L.: What's new under the sun? A corpus linguistic analysis of the 2022 Italian election campaign themes in party manifestos. Modern Italy 29(1), 51–72 (2024). https://doi.org/10.1017/mit.2023.45

contentious issues like the Russian–Ukrainian war. No link to the dataset was provided. Giglietto et al.[5] developed a workflow to detect and monitor coordinated social media accounts, applied to the 2022 Italian election. This workflow built on prior research on coordinated inauthentic behavior from earlier elections. Initial data included 435 coordinated accounts across Facebook and Instagram, with continuous monitoring from July to September 2022. The process identified over 1,000 highly engaged political posts and 272 coordinated links, along with 66 new political accounts and 554 generic coordinated accounts. The study showcased the importance of real-time monitoring to capture evolving tactics in information operations, emphasizing the dynamic nature of coordinated inauthentic behavior on social media.

With respect to these datasets, and the objective of RightNets, ITA-ELECTION-2022 seems to be suited for re-use, to the point that we plan to use those data for the analysis of the 2022 Italian political elections and, instead, to implement our collection for the 2024 European elections.

Data sources and data collection method: the implementation of dedicated scrapers

The data collection plan for the RightNets project involves developing two separate scrapers using Python to gather relevant data from Instagram and X (Twitter). We choose Instagram and X as primary data sources for RightNets due to their significant influence in digital campaigning and political communication. X serves as a crucial platform for public discourse, allowing politicians, public figures, and voters to engage directly, often shaping news agendas and public opinion[6]. Instagram, with its visual-centric approach, is crucial for understanding the impact of images, videos, and stories in political campaigns, especially among younger demographics[7]. Both platforms are highly utilized for micro-targeting and have been pivotal in previous election interferences, making them critical to monitor for foreign influence, covert financing, and compliance with legal standards, thereby aligning perfectly with the objectives of RightNets.

Given that data from social media (including X and Instagram) about the Italian political election in 2022 is available in open access (in "ITA-ELECTION-2022", as explained in the previous section), our scrapers will focus on posts related to the 2024 European elections, specifically targeting posts in Italian containing a selection of keywords related to such ballot (e.g., the hashtags #EUElections2024 or #ElezioniEuropee2024). This approach will allow us to analyze the social media landscape and understand the impact of digital campaigning on political discourse. Below, we detail the implementation of the scrapers for each platform. With the help of the jurists of the team, we will restrict the scraping into a specific time window (e.g., a month) before the date of the election (8-9 June 2024).

Both scrapers will run in a dedicated Docker container, to ensure a clean environment for the script execution as well as the reproducibility of the data collection process.

Instagram scraper

To collect data from Instagram, we plan to use Instaloader[8], a Python library that allows for downloading posts and metadata from public Instagram profiles. The scraper will be designed to target posts containing the specified hashtags, ensuring that the content is relevant to our research focus on the 2024 European Elections. The process involves the following steps:

1. The script initializes Instaloader and sets it to search for posts with the hashtags #EUElections2024 or #ElezioniEuropee2024. The search is confined to posts in Italian to maintain relevance.

[5] Giglietto, F., Marino, G., Mincigrucci, R., Stanziano, A.: A workflow to detect, monitor, and update lists of coordinated social media accounts across time: The case of the 2022 Italian election. Social Media + Society 9(3), 20563051231196866 (2023). https://doi.org/10.1177/20563051231196866

[6] Jungherr, A.: Twitter use in election campaigns: A systematic literature review. Journal of Information Technology & Politics 13, 72–91 (2016). https://doi.org/10.1080/19331681.2015.1132401

[7] Larsson, A.O.: The rise of Instagram as a tool for political communication: A longitudinal study of European political parties and their followers. New Media & Society 25, 2744–2762 (2023). https://doi.org/10.1177/14614448211034158

[8] https://instaloader.github.io/

2. For each post identified, the script extracts key metadata, including the post ID, account name, number of likes, number of comments, and geotag (if available). The post's text (caption) is also captured.
3. Each post's image is downloaded and saved in a dedicated folder named "images," using a progressive numerical ID as the file name (e.g., 1.jpg, 2.jpg). This approach facilitates easy cross-referencing with the CSV data.
4. The script saves the extracted metadata in a CSV file. Each row in the CSV file includes the post ID, the corresponding image ID, the account name, the number of likes and comments, the geotag, and the post's text. This structured format will enable efficient data analysis and visualization.
5. To ensure compliance with Meta's guidelines and policies, the script includes error-handling mechanisms that detect and manage restrictions on data access.

Where Instaloader operations should conflict with Meta's policies or not allow to scrape the planned data, we plan to implement a back solution, using the Meta's Content Library APIs[9], a compliant alternative for accessing Instagram data.

X scraper

For data collection from X, we plan to employ Tweepy[10], a Python library that interfaces with the Twitter API v2, allowing us to access tweets containing the targeted keywords. The implementation will mirror the Instagram scraper to maintain consistency in data structure and analysis. The steps are as follows:

1. The script uses Tweepy to authenticate and connect to the Twitter API, setting search criteria as those used for the Instagram posts, filtering post that are in Italian.
2. For each tweet matching the criteria, the script collects metadata, including tweet ID, the account name, the number of likes and retweets, and the tweet's text.
3. Unlike Instagram, Twitter does not provide direct access to images through the API; however, links to media content (if available) are captured in the metadata.
4. The collected data is saved in a CSV file with a structure identical to the Instagram data: each row includes the tweet ID, account name, number of likes and retweets, and the tweet's text. This consistency ensures compatibility across datasets and facilitates combined analysis.
5. The Twitter scraper is designed with robust error-handling features to manage rate limits and API restrictions, ensuring reliable and uninterrupted data collection.

Data Management and Data collection timeline

The data will be processed in compliance with data protection regulations, using public posts and accounts only, ensuring the ethical handling of personal information. The dataset will be made publicly available, at the end of the project, on a public repository in GitHub to adhere to the FAIR (Findable, Accessible, Interoperable, and Reusable) principles. UAIR principles. Using a public repository will make it Findable, Accessible and Reusable, while using .csv text file to store the data will make it easy interoperable, for data analysis and statics in almost any programming language and data analysis tool. To respect these principles, the source code of the scrapers will be published in the same GitHub open access repository.

Our data collection and public release practices will strictly adhere to the terms of service of the platforms from X and Instagram. We will share raw post, but, instead, we will put in the repository post IDs and URLs, following the example of the cited "ITA-ELECTION-2022." Such IDs and URLs allows the retrieval of the original content, except in cases where posts have been deleted or made private. This approach ensures respect for user privacy and platform rules.

Data collection timeline

The timeline of the data collection is composed of the following phases. It begins with the setup and preparation phase in months 11-12 (October 24 – November 24), where scrapers are developed and tested.

[9] https://transparency.meta.com/it-it/researchtools/meta-content-library/
[10] https://www.tweepy.org/

This is followed by a pilot data collection phase in month 13 (December 24) to evaluate the effectiveness of the scrapers, making refinements if necessary. The main data collection occurs in months 14-15 (January 25, February 25), targeting posts within a defined time window before the June 2024 European elections. In months 16-17 (March 25 – April 25), the collected data is validated and integrated into the broader RightNets project to meet research objectives. Finally, in months 22-23 (September 25 - October 25), the data and scraper source code are prepared for public release, and in month 24 (December 25), the data is published on GitHub, ensuring compliance with ethical standards and accessibility principles.

Months	**Activities**
M11-M12 (Oct 24 – Nov 24)	Setup and preparation: development and testing of the scrapers.
M13 (Dec 24)	Pilot data collection and evaluation of collected data with refinement, if necessary, of the scrapers.
M14-M15 (Jan 25 – Feb 25)	Main data collection for posts about the 2024 European Election, in the time window planned with the jurists, before the 8-9 June 2024.
M16-M17 (Mar 25 – Apr 25)	Validation of collected data with respect to the expected outcomes and integration of the data in the broader RightNets perspective, for the project research objectives.
M22-M23 (Sep 25 – Oct 25)	Preparation of the collected data (and the source code of the scrapers) for the public release (including the creation of file with the post IDs and URLs only and the documentation for the public repository in GitHub).
M24 (Nov 25)	Data release and reporting, with the publication in the GitHub repository.

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

Il Progetto *Normative and Digital Solutions to Counter Threats during National Election Campaigns – RightNets* è finanziato dall'Unione europea - Next Generation EU, nell'ambito del bando PRIN 2022 PNRR Missione 4, Componente 2, Investimento 1.1 (P2022MCYCK; CUP D53D23022340001). I punti di vista e le opinioni espresse sono solo quelli degli Autori e non riflettono necessariamente quelli dell'Unione europea o della Commissione europea. Né l'Unione europea né la Commissione europea possono essere ritenute responsabili per essi.