Finanziato dall'Unione europea NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

Finanziato nell'ambito del Piano Nazionale di Ripresa e Resilienza PNRR. Missione 4, Componente 2, Investimento 1.3 Creazione di "Partenariati estesi alle università, ai centri di ricerca, alle aziende per il finanziamento di progetti di ricerca di base"

GRINS FOUNDATION

SPOKE 2, DELIVERABLE 3.2

"A web-based tools to provide citizens with politicians' statement and related indexes on verifiability and other dimensions"

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

Document data	
Title	Spoke 2 Work Package 3 D3.2 "A web-based tools to provide citizens with politicians' statement and related indexes on verifiability and other dimensions"
Owner	Università di Bologna
Contributors	Università degli Studi di Roma Tor Vergata, Università degli Studi di Padova, Università degli Studi di Torino
Document version	D3.2 – v.2
Last version date	28/02/2025

Executive summary

This document describes the data on politicians statements which underpins a series of web-based tools available on the AMELIA platform. Through these tools, the public can gain easy access to a series of indicators on politicians statements, organized by geographical (Electoral District, Province or Provincial Capital) and temporal (Years of Months) units. The data consists of three blocs, described respectively in Sections 1, 2 and 3 of this document.

The **first bloc of data**, on **"Verifiability Index"**, is contained in attached Stata dataset "verifiability_index_XVIII_legislatura_ProvinceMonthLevel.csv". It was produced by the team coordinated by Prof. Francesco Sobbrio at Università degli Studi di Roma Tor Vergata (and including Drs Francesco Barilari and Abhijay Pandita). It contains two indicators for the extent to which the statements made by top Italian politicians are objectively verifiable (one for politicians belonging to the Senate, and one those belonging to the House of Representatives). The indicators are measured monthly at the relevant electoral district level (Senate or House of Representives), and cover the entire period of the 18th Legislature (April 2018-October 2022). Overall, this dataset provides a novel perspective on the extent to which Italian national politicians use an objective language in their communication strategy.

Web-based tool: Interactive Map of verifiability index by Province-Month

https://ameliadp.grins.it:51904/superset/dashboard/p/edLoP0Ro3Nj/ or
https://ameliadp.grins.it:51904/superset/dashboard/46/

Web-based tool: Timeline of Verifiability Index by Province-Month

https://ameliadp.grins.it:51904/superset/dashboard/p/PVX7zKloKRG/ or
https://ameliadp.grins.it:51904/superset/dashboard/47/

The **second bloc of data**, on **"Local Politicians on Facebook – Topic Analysis"** is contained in the attached csv file "mainTopics_localPoliticians_sm_prov.csv". It was produced by the team coordinated by Prof. Roberto Bonfatti at the Università degli Studi di Padova (and including Drs Duccio Gamannossi degl'Innocenti and Peng Ge, and Professors Edoardo Grillo and Orestis Troumpounis). Over the course of almost two years, the team has downloaded from the Crowdtangle platform over 800,000 Facebook posts from the two mayoral candidates who received the most votes in 1303 large Italian municipalities, over the period 2019-2023. It has then conducted text analysis on these posts, and constructed indicators for the relative importance of fourteen different topics of discussions: the six related to the activities identified by Lgs No. 216 of November 26, 2010 as the main activities carried out by Italian municipalities, plus eight additional topics which recur frequently in the overall sample. The indicators are provided at the province and year level. Overall, this dataset provides a novel opportunity to learn about what Italian local politicians have been talking about the last few years, and thus also a perspective on what are the most important topics for Italian voters.

Web-based tool: Interactive Map of Main Topics Discussed by Local Politicians on Social Media by Province

https://ameliadp.grins.it:51904/superset/dashboard/p/LkjaWBzm2l5/ or
https://ameliadp.grins.it:51904/superset/dashboard/38/

Web-based tool: Interactive Map and Table of Main Topics Discussed by Local Politicians on Social Media by Province

https://ameliadp.grins.it:51904/superset/dashboard/p/edLoPJ8a3Nj/ or
https://ameliadp.grins.it:51904/superset/dashboard/42/

The **last bloc of data**, on **"Sentiment analysis of news about local elections"** was produced by a team coordinated by Prof. Pierluigi Conzo at the Università degli Studi di Torino (and including Dr. Marina Rizzi), and is contained in the attached csv file "sentiment_dataset_province_elezioni_comunali_2013-2024-con_codici_ISTAT". The team has collected all newspaper articles available on Lexis-Nexis and citing the top two candidates in mayoral elections in all Provincial Capitals in Italy, during all election years in 2013-2024. Using state-of-the-art tools of sentiment analysis, the team has then constructed several indicators of the mean and standard deviation of positive and negative sentiment detected in such articles. Such indicators represent a unique opportunity to learn about media-based political polarisation at a local level, for all the above-mentioned municipalities and years.

Web-based tool: Interactive Sentiment Map of News Coverage on Local Elections
https://ameliadp.grins.it:51904/superset/dashboard/52/ or
https://ameliadp.grins.it:51904/superset/dashboard/p/PVX7zXM5mKR/

Web-based tool: Interactive Sentiment Map and Table of News Coverage on Local Elections
https://ameliadp.grins.it:51904/superset/dashboard/56/ or
https://ameliadp.grins.it:51904/superset/dashboard/p/gEWmMLXbo4Y/

TABLE OF CONTENTS

1. Verifiability index

1.1 Short description

The dataset contains two indicators of transparency of politicians' statements. One referred to politicians belonging to the lower chamber (House of Representatives), and one to the one belonging to the Senate. The indicators refer to the share of verifiable statements made by politicians in each month (i.e., total verifiable statements divided by total statements). Both indicators are at the district-year-month level. The dataset reports missing values whenever there are no politicians in the baseline dataset belonging to a given electoral district or when none of the politicians in a given district made a statement in a given year-month.

1.2 Technical details

The dataset is constructed starting from a list of the most relevant politicians active in the XVIII Italian legislature. This list encompasses the prime minister, the deputy prime minister, the Secretary of the Council of Ministers, the party leaders, the president and vice-president of the senate, the president and vice-president of the house, the ministers, the vice-ministers, the party speaker of the house and the party-speaker of the senate.

The set of statements has been obtained by searching news containing the last name of any of the politicians in the title or in the "snippet" of any article published by the main Italian news agencies (ANSA, AGI, Adnkronos, Askanews) between April 2018 and October 2022 (XVIII Italian legislature). The statements are then extracted from the entire set of articles, and each statement is matched with one of the politicians contained in the politicians' list. Finally, each statement is classified as "verifiable" or "non-verifiable" by employing a Support Vector Machine (SVM), a supervised learning model from the field of machine learning. The SVM was trained on a dataset of around twelve thousand politicians' statements that have been manually categorized as "verifiable" or not.

1.3 Codebook

The dataset includes the following variables:

<u>Information on Provincial Identifiers:</u>
The variables "Codice dell'Unità territoriale sovracomunale (valida a fini statistici)," "Codice Regione," "Ripartizione geografica," "Ripartizione geografica," "Denominazione dell'Unità territoriale sovracomunale (valida a fini statistici)," and "Denominazione Regione" serve as provincial identifiers and are sourced from ISTAT (Elenco codici statistici e denominazioni delle unità territoriali; June 30, 2024). "Province_code" was created from the "Sigla Automobilistica" in the same dataset. Together, these variables ensure that each unit of observation is uniquely identified.

Columns Description :

Name of Variable	Description
Codice Regione	ISTAT code that uniquely identifies the region
Codice Ripartizione Geografica	ISTAT code for the Macro Region (i.e., North-West, North-East, Center, South, Islands)
Denominazione dell'Unità territoriale sovracomunale (valida a fini statistici)	ISTAT name for the territorial unit above the municipality level (e.g., province)
CodiceProvinciaStorico1	Historical Province code
Denominazione Regione	ISTAT name for the region
Province_code	ISO 3166-2:IT code for the province
Ripartizione geografica	ISTAT name for the Macro Region (i.e., North-West, North-East, Center, South, Islands)
Circoscrizione_camera	name for the Lower Chamber Italian "circoscrizione" (based on XVIII legislature)
Circoscrizione_senato	name for the Senate Italian "circoscrizione" (based on XVIII legislature)
Share_verifiable	Share of verifiable statements made by politicians in each month (i.e., total verifiable statements divided by total statements).
type	Variable specifying the two Italian branches to which the index referred to: Lower Chamber or the Senate ("respectively camera" and "senato"
Yearmonth	variable for the observation year-month (values in format Year-Month from 2018-04 to 2022-10).
Year_	variable for the observation year (2018-2022)
Month_	variable for the observation month (1-12)

2. Local Politicians on Facebook – Topic Analysis

2.1 Short description

This dataset explores the topics discussed by Italian local politicians on their public Facebook pages. It focuses on the two mayoral candidates who received the most votes in elections between 2019 and 2023. The data includes over 800,000 posts from 2,116 politicians in 1303 municipalities, providing a detailed analysis of the themes covered in their social media activity.

The dataset captures the share of posts on fourteen specific topic categories. The categories used are the six identified by D. Lgs No. 216 of November 26, 2010 (instituting Fabbisogni Standard): General Services and Administration, Security and Local Police, Education, Mobility and Urban Planning, Waste management, and Social Services and Kindergartens. Eight additional categories were added to capture the broader scope of social media communication: COVID-19, Cultural Services and Events, Economic Activities, Festivities and Public Events, Politics, Sport, Utilities, and Other.

To ensure privacy and anonymity, the dataset includes only aggregated data at the provincial level. Provinces are eligible if they meet the following criteria: 1) At least three politicians posting per year and, 2) a minimum of 20 posts per year. Table 1 provides summary statistics of the dataset.

Table 1: Summary Statistics of Non-Empty Messaged Posts by Eligible Provinces

Year	Total Posts	Unique Politicians	Mean Posts by Politician	Median Posts by Politician	Unique Provinces	Mean Posts by Province	Median Posts by Province
2019	99787	350	285	158	52	1919	1165
2020	132876	379	351	214	54	2461	1649
2021	124211	441	282	152	57	2179	1198
2022	219190	1699	129	67	86	2549	1524
2023	256135	1912	134	68	87	2944	1882
Whole Sample	832199	2116	393	142	88	9457	5210

2.2 Technical details

The analysis focuses on municipalities ranked among the top 2,500 by population in 2020, with a minimum population of 3,856 (e.g., Romagnano Sesia, Novara) for the period 2022–2023. For the years 2019–2021, it includes the top 1,363 municipalities by population in 2020, with a minimum population of 7,821 (e.g., Montorio al Vomano, Teramo).

Topic identification involved both automatic and manual steps. The automatic step utilized BERTopic (Grootendorst, 2022) with the multilingual embedding model "paraphrase-multilingual-mpnet-base-v2" and grid search to optimize hyperparameters for relative validity of the HBSCAN clustering model (McInnes et al., 2017). This process classified posts into 139 topics. The manual step grouped these topics into the 14 categories outlined in the Codebook. The cross table between the 139 topics and the fourteen categories is available from the authors upon request.

2.3 Codebook

The dataset includes the following variables:

Information on Provincial Identifiers:
The variables "Codice dell'Unità territoriale sovracomunale (valida a fini statistici)," "Codice Regione," "Ripartizione geografica," "Ripartizione geografica," "Denominazione dell'Unità territoriale sovracomunale (valida a fini statistici)," and "Denominazione Regione" serve as provincial identifiers and are sourced from ISTAT (Elenco codici statistici e denominazioni delle unità territoriali; June 30, 2024). "Province_code" was created from the "Sigla Automobilistica" in the same dataset. Together, these variables ensure that each unit of observation is uniquely identified.

Columns Description :

Name of Variable	Description
Codice dell'Unità territoriale sovracomunale (valida a fini statistici)	ISTAT code for the territorial unit above the municipality level (i.e., province)
Codice Regione	ISTAT code that uniquely identifies the region
Codice Ripartizione Geografica	ISTAT code for the Macro Region (i.e., North-West, North-East, Center, South, Islands)
Denominazione dell'Unità territoriale sovracomunale (valida a fini statistici)	ISTAT name for the territorial unit above the municipality level (e.g., province)

Denominazione Regione	ISTAT name for the region
Province_code	ISO 3166-2:IT code for the province
Ripartizione geografica	ISTAT name for the Macro Region (i.e., North-West, North-East, Center, South, Islands)
value	Share of Facebook posts by local politicians that fall into a given topic category
var	Variable specifying the topic category of social media posts by local politicians (values in 'Covid', 'Cultural Services and Events', 'Education', 'Festivities and Public Events', 'General Services and Administration', 'Local Economy', 'Mobility and Territory', 'Other', 'Politics', 'Security and Local Police', 'Social Services and Kindergardens', 'Sport', 'Utilities', 'Waste Management' Economy, Healthcare, Infrastructure)
year	variable for the observation year (values in 2019-2023)

3. Sentiment Analysis of news about local elections

3.1 Short Description

The dataset contains information about the sentiment used in articles that talked about the first two candidates in local elections for provincial capitals from 2013 to 2024. Sentiment information are available for all the newspapers in the datasets, and also for ANSA (the press agency's publication) and non ANSA newspapers. The dataset presents missing values if the provincial capital has not undergone a local election in that year.

3.2 Technical Details

The dataset has been constructed starting from the news database available on Nexis Lexis. For each local election of provincial capitals that was held between 2013 and 2024, we selected the first two candidates. For each candidate, we then selected all the articles in Nexis Lexis that mention the surname of the candidate and the name of the municipality during the year of the local election.

We then use a state-of-the-art algorithm to perform sentiment analysis on Italian text, FEEL-IT, an instrument made available by the MilaNLP Lab at Bocconi University. For each article, FEEL-IT provide a score between 0 and 1 indicating how positive and how negative the article is. We then obtain a measure of positive sentiment and a measure of negative sentiment for each news in our database. We then collapse these measures at the municipality level. In particular, for each municipality in each election year, we calculate the mean and the standard deviation of the positive sentiment, and separately of the negative sentiment, so that, for each municipality and each year we obtain two summary measures of the positive and negative sentiment of the news that referred to one of the two candidates in the local elections of that municipality. We calculated these metrics for all the news, then only for ANSA news, and finally only for non-ANSA news.

3.3 Codebook

The dataset includes the following variables:

Information on Provincial Identifiers:
The variables "Codice dell'Unità territoriale sovracomunale (valida a fini statistici)," "Codice Regione," "Ripartizione geografica," "Ripartizione geografica," "Denominazione dell'Unità

territoriale sovracomunale (valida a fini statistici)," and "Denominazione Regione" serve as provincial identifiers and are sourced from ISTAT (Elenco codici statistici e denominazioni delle unità territoriali; June 30, 2024). "Province_code" was created from the "Sigla Automobilistica" in the same dataset. Together, these variables ensure that each unit of observation is uniquely identified.

Columns Description :

Name of Variable	Description
Codice dell'Unità territoriale sovracomunale (valida a fini statistici)	ISTAT code for the territorial unit above the municipality level (i.e., province)
Codice Regione	ISTAT code that uniquely identifies the region
Codice Ripartizione Geografica	ISTAT code for the Macro Region (i.e., North-West, North-East, Center, South, Islands)
Denominazione dell'Unità territoriale sovracomunale (valida a fini statistici)	ISTAT name for the territorial unit above the municipality level (e.g., province)
Denominazione Regione	ISTAT name for the region
Province_code	ISO 3166-2:IT code for the province
Ripartizione geografica	ISTAT name for the Macro Region (i.e., North-West, North-East, Center, South, Islands)
score	Score of the measure of the predicted sentiment that falls into that newspaper category
var	Variable specifying the measure of the sentiment category considered and the type of newspapers considered (values in 'Avg_Positive_Sentiment', 'Std_Positive_Sentiment', 'Avg_Negative_Sentiment', 'Std_Negative_Sentiment', 'Avg_Positive_Sentiment_ANSA', 'Std_Positive_Sentiment_ANSA', 'Avg_Negative_Sentiment_ANSA', 'Std_Negative_Sentiment_ANSA', 'Avg_Positive_Sentiment_NOT_ANSA', 'Std_Positive_Sentiment_NOT_ANSA', 'Avg_Negative_Sentiment_NOT_ANSA', 'Std_Negative_Sentiment_NOT_ANSA', 'Difference_positive_sent_Non_ANSA_VS_ANSA', 'Difference_negative_sent_Non_ANSA_VS_ANSA')