THE ONLINE SERVICES OF MUNICIPALITIES IN ITALY: THE DIGITAL DIVIDE [1]

Chiara Orsini, Fabrizio De Fausti, Sergio Leonardi

Abstract. In Italy, the enhancement of online public services is one of the main goals of the National Recovery and Resilience Plan (NRRP), including relevant investment on digital identity. To this respect, traditional methods for data collection in official statistics may lack the capability to grasp relevant information to assess the digital transition of public institutions, especially in a timely fashion.
The main objective of the paper is to measure in a systematic manner the capacity of local institutions in offering online public services, by collecting relevant information directly through municipal websites using Machine Learning (ML) techniques, while highlighting territorial digital gaps in providing online services. Particularly, the study aims to develop an automatic classification framework to investigate whether and to what extent Italian municipalities implement the digital identity system. This is achieved by comparing the effectiveness of random forest and naive Bayes supervised ML algorithms commonly used for text classification. The classification procedure is based on two different approaches: 1) the integration and use of auxiliary online sources with official statistics sources, such as the Permanent Census of Public Institutions conducted by the Italian National Institute of Statistics (ISTAT) in 2023, and 2) gathering information on relevant features of municipalities' websites by means of web scraping techniques.
By combining official statistics information with big data, the analysis draws the attention on municipalities' digital divide, by comparing online access to public services of citizens living in different areas of the Country, e.g. regions and provinces.

1. The digitalization of public services

Digitalization is a cornerstone of the EU's strategy for future growth and development. It enhances economic competitiveness, improves public services, fosters social inclusion, and supports sustainable development. Particularly, the

[1] This Working Paper is the result of the joint work of Chiara Orsini, Fabrizio De Fausti, Sergio Leonardi. However, each chapter was drafted by a lead author. The lead author of the paragraphs 1, 4, 5 is Chiara Orsini, the paragraphs 2 is drafted by Sergio Leonardi and the paragraphs 3 and 5 by Fabrizio De Fausti.
.

"Digital Decade"[2], the initiative by the European Union aimed at empowering Europe's digital transformation by 2030, focuses on enhancing the EU's digital capabilities and infrastructure, by ensuring that Europe remains competitive and sovereign in the digital age. The comprehensive strategy is intended to harness the potential of digital technologies, making Europe a leader in the digital world while ensuring that this transformation benefits all citizens and businesses in a sustainable, secure, and inclusive manner.

The digitalization of public services[3] is one of the main priorities in the EU, consisting of a multifaceted effort involving significant investments, policy initiatives, and technological advancements. While substantial progress has been made, ongoing efforts are needed to address infrastructure gaps, improve digital literacy, and ensure the security and efficiency of digital public services. By continuing to focus on these areas, the EU creates a structured, transparent and shared monitoring system based on the Digital Economy and Society Index (DESI) to measure progress towards human capital, connectivity, integration of digital technology and digital public services. In Italy the digitalization of public services is a significant aspect of Italian NRRP[4], which has invested 255 million euro in the use of digital identity. This process aims to improve the efficiency, accessibility, and transparency of Public Administration (PA) and services, by enhancing digital public services, modernizing infrastructure, and promoting digital literacy, thorough using digital identity systems[5].

In Italy measuring digitalization of municipalities is essential for driving effective digital transformation. It helps in assessing progress, guiding resource allocation, identifying best practices, enhancing public services, ensuring transparency, and fostering economic and social development. Regularly evaluating performance of municipalities progress towards the online access of public services is needed. The main objective of the paper is to provide a new methodological model to support official statistics in measuring, in a systematic manner how and to what extent the public sector is conducting the digitalization transition of local institutions. Particularly, the ML techniques and web scraping methods are studied and tested in this paper to investigate the capacity of local institutions in offering online public services, highlighting territorial digital gaps. Moreover, the study is intended to show the opportunities and challenges when ML methods are combined

[2] European Commission, Communication establishing the Union-level projected trajectories for the digital targets, 2023.

[3] European Commission, Digital Economy and Society Index Report— Digital Public Services, 2022.

[4] Presidenza del Consiglio dei ministri, Piano Nazionale di Ripresa e resilienza, 2021.

[5] Digital identity system consists of the use of SPID (Public Digital Identity System) or CIE (Electronic Identity Card), or the National Service Card (Carta Nazionale dei Servizi - CNS) a unified digital identity that allows citizens and businesses to access online services securely.

with official statistics. The digitalization of local institutions is conducted by identifying inhabitants who have online access through digital identity in the municipalities, measured as the share of municipalities having SPID/CNS/CIE gateway on their websites, as well as inhabitants who have online access (single gateway for construction), measured as the share of municipalities implementing SUE (Sportello Unico per l'Edilizia) on their websites. Furthermore, the methodological approach consists of using official statistics to have the list of municipalities (7,904) and their structural characteristics through the 2022 Permanent Census of the Public Institutions[6]. The list of municipalities has been processed through web scraping technique, applied as data collection in creating the Term Document Matrix (TDM), while creating a labelled dataset, though python routine, to identify a municipality sample which identify the institutions which implemented digital identity and SUE as online services. Moreover, the study tests several supervised ML algorithms to make inference on Italian municipalities, which use digital identity system and make SUE as online service. The replicable model is intended to offer updated data of digital transition of local public institutions. Previous studies have demonstrated the effectiveness of extracting information from websites to enrich and automate processes that support statistical analyses using machine learning (De Fausti *et al.* 2019). This study aims to explore for the first time the potential of machine learning algorithms to automate and enrich information on the digitalization of administrations at the local level.

2. Web scraping technique

Generic web scraping is a technique that aims to extract information from municipalities' websites leading to a TDM, which summarizes in an incidence matrix such information. Considering the number of municipalities, 7,904 at December 31, 2022, and the pages per municipality website, potentially unlimited, two main phases have been processed: information retrieval and storage, matrix construction.

In the first phase, municipalities' Uniform Resource Locator (URL) list has been cleaned from transcription mistakes and web protocol updates. Afterwards, RootJuice application was applied to visit the cleaned list of sites and extract relevant information from it. Table 1 outlines the main results of web scraping conducted by RootJuice.

[6] The Permanent Census of Public Institutions is conducted by the Italian Institute of Statistics every three years, combining administrative sources and information gathered through surveys. Istat, Permanent Census of Public Institutions, 2022.

Table 1 – *RootJuice report for scraping websites.*

MAIN ITEMS	RESULTS
Total visited pages	68,562
Total municipality IDS detected	7,991
Number of municipalities with at least 1 page retrieved	6,239
Number of municipalities with 0 page retrieved	1,662
Number of URLs in seed file	7,903
Number of URLs filtered out	2
Number of URLs after filter	7,991
Reached sites percentage	79,0
Started and ending data time	11/03/2023 time 10:48-22:45

Source: RootJuice output run by the authors

RootJuice makes use of a list of forbidden domains, to avoid visiting non-relevant sites, e.g. government or general-purpose websites. Other configuration parameters allow to increase the visit number of pages that could be visited starting from the root of a municipality site) and the total number of pages per Municipality site. After several RootJuice runs we reach 98% municipalities sites, that is 7,741, missing 162, mainly because of site structural problems such as SSL (Secure Sockets Layer) certificate issues, website unavailable at scraping time, outdated or invalid URL. RootJuice, as per its name, squeezes textual information from each visited page, storing them in a CSV structure, a textual Comma-Separated Values file, in which each column holds text taken from a different page source (e.g. HTML[7] elements, alt properties, title properties and so on) along with their navigation properties (e.g. page URL, referral, URL list element and so on). This huge CSV file is stored in an Apache SolR collection by SolrTsvImporter application, which chunks the file and send it to the Apache SolR collection. To tune this step we configure in SolrTsvImporter the number of threads and the number of CSV rows per chunk. Apache SolR stores all the content, one chunk at a time, indexing all of the words in the collection, thus preparing them for the next phase. At the end of this step it's possible to check the matching between the number of content retrieved from RootJuice and the size of the SolR collection.

In the second phase, we use FirmsDocTermMatrixGenerator to build the TDM from the Apache SolR collection. Application configuration parameters span several aspects of natural language parsing. The first stage is the identification of a primary and a secondary language of the textual content. For the sake of this work, we use just the Italian language. The second stage is the specification of a special set of words to add to the language, to include acronyms, specific language and so on. We call this set go words list and comprehends a few relevant acronyms.

[7] HyperText Markup Language

Another parameter is the set of words to ignore in the analysis because their frequency in the language is too high. We call this set stop words list, which comprehends Italian and English stop words. We use English stop words in order to reduce web programming language noise, whose terms are mostly English words.

We configure a range of word length, in order to filter unwanted noisy text. In our runs we configure a minimum of 3 and a maximum of 25 letters for a valid word. The application creates the column list from SolR indexed words by removing words whose length is outside of the range and removing words TreeTagger does not recognize to be in the Italian dictionary or by go words list, then removing words from stop words list. For each remaining word, we obtain its stem, with Snowball Stemmer, and put this stem in the column index of the TDM.

The application counts the number of occurrences of each stem in each content, aggregating them on a per-municipality basis, leading to an incidence matrix with one row per municipality and one column per stem. Table 2 outlines the main configurations of the TDM.

Table 2 – *Parameters in creating the TDM.*

MAIN ITEMS	RESULTS
Number of rows (Municipalities)	6,239
Number of columns (words)	12,807
Starting data time	19/03/2023 07:20
Ending data time	19/03/2023 07:50

Source: TDM processing run by the author

3. ML models

The textual data collected from Italian municipal websites using web scraping techniques described in section 2 aims at extracting information and keywords from the websites, focusing on the digital services offered by each municipality. Also, in the first phase of data collection and pre-processing, the cleaned textual data is tokenized and processed through stemming. Then, the data are gathered into a TDM, which is a fundamental data structure for the analysis, and the ML algorithms are used to classify and identify the presence or absence of online services for the citizens living in a certain municipality. TDM represents a sparse matrix with the frequency of terms from a dictionary existing in the documents (municipal websites). This matrix facilitates the application of ML algorithms by quantifying the textual data in a form that the algorithms can process.

In the test, two separate classifiers are trained: one is trained in order to identify the presence or absence of the SPID identification service, and the other in order to

identify the SUE online service. The target variable is therefore dichotomous and indicates (1 = presence of the service, 0 = absence of the service).

The SPID target variable is automatically labelled using the technique described in section 3. Whereas the SUE target variable requires manual labelling processing conducted on 500 municipalities, which were randomly sampled from the total of 7,904 municipalities gathered by the TDM.

In preparing the model training, the TDM is re-designed by excluding documents that contain insufficient information. Particularly, municipalities with only three or fewer terms were omitted, reducing the dataset to 448 municipalities.

For feature selection, the chi-squared (χ^2) statistic was used to identify terms that are statistically significant with respect to the class labels. Various models are trained, by using terms that exceed the 70th, 80th, 90th, 93rd, 95th, and 99th percentiles in chi-squared values. The performance is evaluated using accuracy metrics and the average F1-macro score, calculated on the test dataset derived from a 70%-30% train-test split of the dataset. This division is commonly adopted for dataset splitting (Kohavi 1995). To enhance the robustness of the model validation in terms of mean and variance, a bootstrap approach (Efron, and Tibshirani 1994) is implemented, conducting 10 different runs.

The study focuses on training and comparing three different ML models widely recognized for their effectiveness in text classification. This methodological choice aims to assess which model best suits the characteristics of the dataset and it provides the most reliable performance in recognizing and in classifying the digital services offered by Italian municipalities. The models selected for this comparison include multinomial naive Bayes, Bernoulli naive Bayes, and random forest, each of which has distinctive features that could influence the accuracy and effectiveness of classification in different contexts.

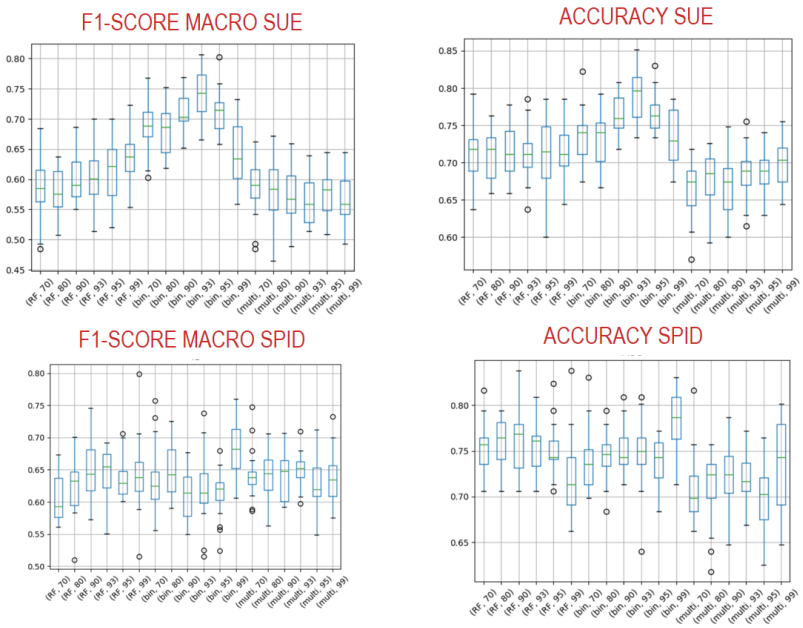
Multinomial naive Bayes classifier (Mccallum and Nigam 1998) is particularly suited for classification with discrete features (e.g., word counts for text classification). It calculates the probability of each category based on the frequency of words and the Bayes theorem, with the assumption that features are conditionally independent given the class label. It is known for its simplicity and effectiveness in handling large datasets.

Bernoulli naive Bayes represents a variant of the multinomial naive Bayes approach. Distinctly, this model utilizes binary feature representation to denote the presence or absence of terms, rather than their frequency counts, it is particularly adept at handling datasets where the mere occurrence of specific terms is more telling than their frequency.

Random forest (Breiman 2001) is an ensemble learning method used for classification, regression, and other tasks that operates by constructing a multitude of decision trees during the training phase. Each tree is trained on a random subset

of the training data, obtained through bootstrap sampling. Through the bootstrap mechanism and the aggregation of each tree's decision—known as bagging—this approach allows each tree to learn from a slightly different version of the dataset, thereby reducing the overall model variance without increasing bias. This makes the model less sensitive to outliers and erroneous data, enhancing its robustness against noise and its ability to manage overfitting.

Figure 1 - *Boxplot of F1- macro score and accuracy over 10 runs for each ML model (random-forest:RF, multinomial-naive-bayes:multi and Bernoulli-naive bayes:bin) and chi-square cut-off for SUE and SPID classifier.*



Source: ML processing run by the author

For each model and for each feature selection made by cutting off at percentiles relative to the chi-squared statistic, we perform 10 runs to account for variability. Figure 1 displays the boxplots pertaining to the performance of the SUE classifier and the SPID classifier.

With regards to the SUE classifier, the choice of the best model is easy; indeed, a peak in performance for the Bernoulli naive Bayes model with a cut-off at the 93rd percentile is observed, the measure of macro F1-score is 74.1% and the accuracy is 79.1%.

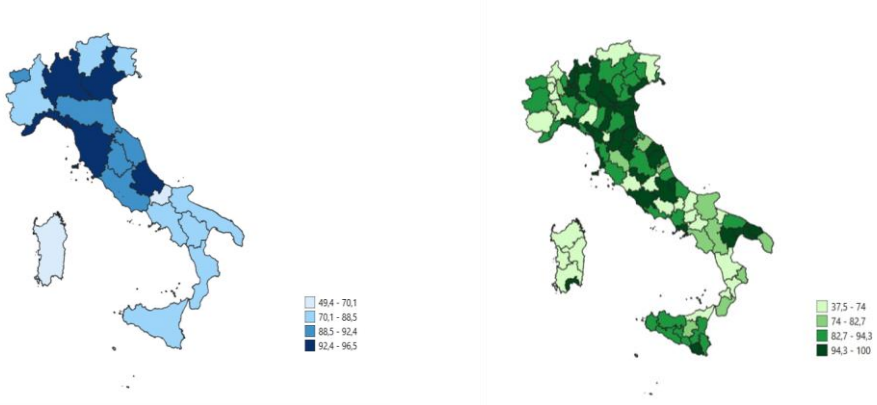
Concerning the SPID classifier, the selection of the best model is not as clear-cut. We ultimately chose the random forest model with a chi-squared cut-off at the 90th percentile, which achieves a macro F1-score of 64.8% and an accuracy of 76.3%. Although this is not the model with the highest scores, it has the lowest cut-off, allowing for a larger number of features on which the model can perform classification. This choice enhances the classifier generalization capabilities.

The models selected and trained on the sample of 448 municipalities are then applied to make inferences and automatically classify all Italian municipal websites collected during the web scraping process.

4. Main results: Services provided by Municipalities

According to the data processed by ML model, in Italy 87.7% of the population has access to public services through the identity systems, particularly 9 inhabitants over 10 have access in Liguria, Lombardia and Veneto, as well as by province Forlì-Cesena, Prato, Trieste have registered the same level.

Figure 2 - *Population with access to public services through digital identity by region and province.*

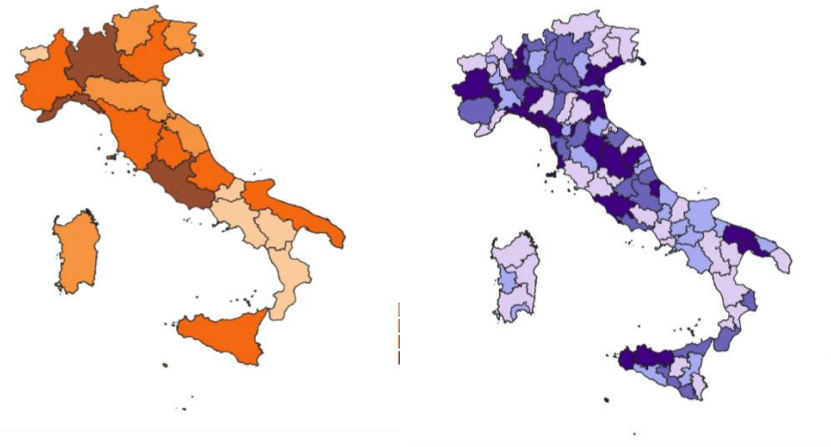


Source: QGIS[8] output run by the author

While in Calabria, Molise and Sardinia less than 7 inhabitants over 10 manage to have access to the identity system and by local areas the municipalities located in Vibo Valentia, Isernia, Oristano show a percentage below 50 per cent. (Figure 2)

[8] QGIS is a geographical information system software

Figure 3 - *Population with access to SUE as online services, by region and province*.



Source: QGIS output run by the author

In Italy 44.4% of the population have online access to SUE, particularly 6 inhabitants over 10 have online access through the municipalities located in Lazio, Lombardia, Liguria and 8 inhabitants over 10 benefit from SUE services provided by the local institutions in Trieste, Genova, Como. Whereas, less than 2 inhabitants over 10 have online access to SUE in Val d'Aosta, Basilicata, Molise and by province in Enna, Matera, Imperia.

The integration of official statistics data lead to profile the municipalities which implemented the digital identity system and online SUE service, particularly the municipalities with digital identity system are mainly located in urban areas in the northeast, employing more than 250 workers and with 20,000 inhabitants and over, while the municipalities with SUE are mainly located in Urban areas in the northwest, employing more than 250 workers and with 20,000 inhabitants and over. Those profiles are in line with official statistics data disseminated by survey on information and communication technologies in the PA[9].

[9] The survey aims to acquire information on the technological equipment available in the Administrations to support internal administrative activities and relations with citizens, businesses and other PAs, in compliance with EU Regulation.

5. Conclusions and lessons learned

The paper shows how official statistics and machine learning methods may be combined to investigate digitalization at the local level in Italy. The estimation produced by using ML models are fully in line with the official statistics results on public institutions digitalization disseminated by Istat in 2023. Particularly, the study shows that the citizens living in urban areas and in the most populated municipalities are more likely to have access online to public services than people living in remote areas. On the other hand, the analysis highlights that the small size Municipalities located in the North-east of Italy manages to provide online services through centralising ICT provider, which is their Region website.

From a methodological point of view, the study explores the application of ML techniques to analyze and classify the digital services provided by Italian municipalities, utilizing data extracted via web scraping. By employing a TDM and three distinctive ML models—multinomial naive Bayes, Bernoulli naive Bayes, and random forest—the analysis highlights the performance of each model.

This research introduces a new approach which adds timeliness and detailed territorial insights to the analysis. The integration of chi-squared for feature selection is useful, in order to trim the inputs for ML models while obtaining more accurate classifications.

The findings provide municipalities with insights of digital service provisions. Effective classification and analysis of digital services may support policymakers to identify service gaps and to plan policy adjustment, in order to have efficient resource allocation and fostering digitally inclusive governance.

Further research is needed to explore the integration of advanced natural language processing techniques to handle linguistic issues more effectively. Expanding the dataset to include a broader range of municipal profiles is needed to enhance the robustness and applicability of the models. Moreover, incorporating more sophisticated ML algorithms such as transformers and BERT (Bidirectional Encoder Representations from Transformers) may provide new opportunities to strengthen the classification accuracy and processing efficiency.

Moreover, the creation of a systematic approach in gathering data through web scraping and measuring digitalization can be applied on other statistical units: enterprises and nonprofit institutions, for instance.

To conclude, the study is intended to show the application of ML models in public sector analytics, as well as it creates a basis for the development of innovative research for digital governance and data-driven decision-making processes. The approach and the result of the working paper provides significant insights on how digital services can be measured and analysed to monitor digital policy advancements.

References

BREIMAN L. 2001. Random Forests. *Machine Learning*, Vol. 45, No. 1, pp. 5-32.
DE FAUSTI F., PUGLIESE F., ZARDETTO D. 2019. Towards automated website classification by deep learning. arXiv preprint arXiv:1910.0999
DE PANIZZA A., LOMBARDI S. 2024. Rapporto sulle istituzioni pubbliche, Istat, ISBN 978-88-458-2146-2.
EFRON B., TIBSHIRANI R. J. 1994. *An introduction to the bootstrap*. Chapman and Hall/CRC.]
EUROPEAN COMMISSION 2023. *Communication establishing the Union-level projected trajectories for the digital targets*.
ISTAT 2022. Permanent Census of Public Institutions. https://www.istat.it/statistiche-per-temi/censimenti/istituzioni-pubbliche/
ISTAT 2022. Survey on information and communication technologies in the PA.
KOHAVI R. 1995. *A study of cross-validation and bootstrap for accuracy estimation and model selection*. Morgan Kaufman Publishing.
MANNING C. D. 2009. *An introduction to information retrieval.*
MCCALLUM A., NIGAM K. 1998. A comparison of event models for Naive Bayes text classification. In *AAAI-98 workshop on learning for text categorization*.
PORTER M. F., 2001. An algorithm for suffix stripping, *Program*, Vol. 14, No.3, pp. 130-137. http://snowball.tartarus.org/texts/introduction.html
PRESIDENZA DEL CONSIGLIO DEI MINISTRI, 2021. *Piano Nazionale di Ripresa e resilienza.*

Chiara ORSINI, ISTAT, chorsini@istat.it
Fabrizio DE FAUSTI, ISTAT, defausti@istat.it
Sergio LEONARDI, PRESIDENZA DEL CONSIGLIO, s.leonardi@governo.it