AI-supported Certification of Family-Friendly Organizations

Davide Vandelli[1], Sara Tonelli[1], Pietro Marzani[1] and Alessio Palmero Aprosio[2]

[1]Fondazione Bruno Kessler, via Sommarive 18, Trento (Italy)
[2]Department of Psychology and Cognitive Science, University of Trento, Corso Bettini 84, Rovereto (Italy)

Abstract

In the Autonomous Province of Trento, the certification issued by the Agency for Social cohesion to municipalities and companies to recognize their commitment to family-friendly solutions has become more and more popular over the years. However, the application process, which foresees the preparation of plans including several actions classified according to a taxonomy, is rather complex and may benefit from domain knowledge coming from plans already in place. To address this requirement, we have designed an AI-supported platform that assists operators in preparing plans by suggesting information on action categories and on plans submitted by similar organizations. It also provides an analytical tool for the Agency to perform periodic revisions of the taxonomy of actions and suggest changes in the categories. The first version of the tools has been assessed by stakeholders, who have appreciated the integration of NLP-based suggestion tools in the process without altering too much the submission workflow.

Keywords

NLP, deep learning, classification, family friendly, taxonomy, certification, plan of action, public administration

1. Introduction

In the Autonomous Province of Trento, a considerable part of the activities of the *Agency for Social Cohesion*, the structure responsible for the implementation of family policies, concerns the development and dissemination of a system of “family brands”, i.e. certifications that recognize the commitment of public and private organizations in adopting family-friendly solutions for their staff and for the families residing in their territory. To obtain this certification, municipalities and companies (henceforth *organizations*) have to submit to the *Agency for Social Cohesion* a “family plan”, where a list of activities indicates what the organization is committed to for implementing work-life balance or family-friendly measures, depending on their range of action. If the plan is approved, a“family brand” is acknowledged to the organization. Two types of “family brands” are provided: the *Family Audit certification* for companies, and the *Family in Trentino label* for municipalities, thus putting in place services that respond to the needs and expectations expressed by families in the area as efforts from the public and private sector. These initiatives have also begun to spread from Trentino (100 municipalities) to the national level, with 60 municipalities receiving the *Family in Italy* label and approximately 150 organizations certified under the *Family Audit* scheme.

The certification process started in 2008 and enabled the collection of a database of information: the Family Audit plans submitted by companies contain more than $9,000$ work-life balance actions adopted in favor of their staff by 320 companies nationwide. The municipal plans submitted by more than 100 Family certified municipalities in Trentino, contain instead more than $4,000$ actions. Actions are classified based on a taxonomy that is specific to each certification in both types of “family brands”.

Given the increasing success of the “family brands” and the growing number of organizations who would like to obtain the certification at national level, it is important to implement a workflow to create and submit the plans that ensures consistency across different operators and enables taking advantage of

Ital-IA 2025: 5th National Conference on Artificial Intelligence, organized by CINI, June 23-24, 2025, Trieste, Italy

dvandelli@fbk.eu (D. Vandelli); satonelli@fbk.eu (S. Tonelli); pmarzani@fbk.eu (P. Marzani); a.palmeroaprosio@unitn.it (A. Palmero Aprosio)
0009-0008-2610-9916 (D. Vandelli); 0000-0001-8010-6689 (S. Tonelli); 0000-0002-1484-0882 (A. Palmero Aprosio)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2025-12-07

existing knowledge about past plans. Indeed, the plans to obtain *Family Audit certification* and the *Family in Trentino label* are already submitted electronically via a platform (see Section 2), but each operator starts the process from scratch and cannot see what other organizations did. Also, the process is fully manual. In this framework, the project "PNC-A.1.3- Digitalizzazione della pubblica amministrazione della Provincia autonoma di Trento" aims to create an Artificial Intelligence (AI) solution specifically designed to support operators in submitting the plans and to allow the *Agency for Social Cohesion* to monitor the submitted requests.[1] The solution proposes a re-design of the current platforms so that writing and submitting a plan will be supported and guided by an LLM-based dialogue system, which will trigger different NLP components. In the remainder of this paper we will focus mainly on the description of such components which will provide *i)* similarity scores between organizations, *ii)* action classification and *iii)* suggestions for taxonomy modifications (Section 3).

2. Family in Trentino Platforms: Current State

The planning tools that allow the submission of the plans use two separate online management systems: the *GeAPF platform* for the *Family Audit certification* and the *Family Plan platform* for the *Family in Trentino label*. Each organization must prepare a plan that consists of several actions, descriptions and objectives to be achieved, each addressing different aspects of family support. In the case of municipalities (*Family in Trentino*) it is a yearly submittal, while for companies (*Family Audit*) the plan is submitted only at the beginning of the process scheme. When entering new actions in the plans, the compiler is asked to provide a textual description of the actions themselves and to classify each of them within a given taxonomy of actions. For example, an action could be:

> *"L'amministrazione comunale intende introdurre una nuova agevolazione per famiglie numerose nelle tariffe del servizio asili nido, in particolare la riduzione della quota fissa mensile in caso di ammissioni di fratelli o sorelle nella misura del 15% per il terzo fratello ammesso e seguenti"*

In this case, the taxonomy label would be *Agevolazioni specifiche per famiglie numerose*, which belongs to the *Misure economiche* macro-category. This classification plays a crucial role because it allows for aggregations and analyses of the input data. For instance, the Agency for Social Cohesion can monitor the data provided by the various organizations and use their outcomes to design additional policies. The taxonomy also serves as a guide for those responsible for drafting the plan, who can draw inspiration from its categories to design actions to be implemented within their organization.

The presence of a reference taxonomy for family plan actions is a valuable element. However, analysis of the use of this taxonomy shows several limitations. In particular, the independent compilation of plans by organization leaders results in differences of interpretation in the application of the taxonomy's items, also fostered by the very nature of the taxonomy, which is considered a complex and overly populated tool (about 200 entries). In order to be effective, the taxonomy must be a flexible and dynamic tool that adapts to the specific situations in the territories and that is enriched as organizations identify new types of actions. Therefore, the development of tools that can support both the effective use of the taxonomy by individual organizations, as well as the maintenance and evolution of the taxonomy itself over time, is essential.

Annotation according to the given taxonomy is not the only aspect in which plan entry by the organizations' operators can be improved: even the descriptive part is often presented differently by different organizations, both in terms of content structure and of detail and quality of the content itself. An appropriate support tool can qualitatively improve the descriptions provided, while helping operators in the use of the system. Advanced features can also help identify actions of interest to a particular organization, for example by suggesting possible relevant actions developed by other similar organizations.

[1] Other activities foreseen in the project involve the domains of protezione civile and tourism but we do not address them in this paper.

3. AI-Supported Platform

The technological solution developed within the project "PNC-A.1.3- Digitalizzazione della pubblica amministrazione della Provincia autonoma di Trento" foresees the implementation of a platform where some functionalities are enabled when operators log in, while some others are active only for the *Agency of Social Cohesion*. Indeed, the platform supports both the submission of family plans to obtain a *Family Audit* (for public and private organizations) or *Family in Trentino* (for municipalities) certification, and the monitoring of the plans by the *Agency of Social Cohesion*. After login, an operator can start the process to submit a plan and optionally ask to be supported by AI. If this option is selected, a chatbot will be activated, which will provide suggestions and support in writing the plan. The operator can ask to take inspiration from plans submitted by other similar organizations in the past, which activates the tool for the suggestion of similar organizations (Section 3.1). When entering the actions that an operator plans to undertake to obtain the certification, a second component can be triggered suggesting one or more categories to be associated with each action (Section 3.2). A third additional component is available only to the *Agency of Social Cohesion* personnel, which matches the taxonomy with the submitted actions and suggests changes to the taxonomy itself, such as label deletion, insertion or merging (Section 3.3). The three components are detailed below.

3.1. Suggestion of Similar Organizations

While entering a new plan, an organization may want to check what other similar organizations have done in the past and which activities others have proposed. This would provide an opportunity to take a look at actions by other organizations, offering records of replicable successful cases. We implement a component that, given an organization in input and one or more similarity criteria, outputs a ranked list of similar organizations. For municipalities and organizations, similarity can be computed based in the criteria shown in Table 1.

Table 1
Comparison of parameters available for similarity.

	Municipality par.	**Company par.**
Plan data	Actions submitted, Descriptions of actions, Macro-categories, Years of certification	*idem as municip.*
Other features	Urban density, Elevation, Population, Distance from the province's main city	Statistical composition of personnel, Sector of the company

We computed the similarity as the inverse of the scaled pairwise Euclidean distance for numerical data, and the Jaccard similarity coefficient score for categorical data. In the numerical case, we end up assigning a similarity that is inversely proportional to the distance between two organizations – a greater distance results in a proportionally smaller similarity. The distance is computed iterating over all combinations of organizations in $\boldsymbol{O} \in \mathbb{R}^m$ as $\boldsymbol{d} = \{|O_i - O_j| \mid 1 \leq i < j \leq m\}$ where the flat vector $\boldsymbol{d} = (d_k)_{k=1}^{K}$ represents the pair combinations of the organizations considered. Then, $\boldsymbol{d}$ is linearly scaled as $d'_k = [d_k - \min(\boldsymbol{d})]/[\max(\boldsymbol{d}) - \min(\boldsymbol{d})]$ for $k \in 1, ..., K$. Ultimately, the inverse of the scaled distance $\boldsymbol{d} \in [0, 1]$ is turned into an intuitive pairwise similarity by $s_k = 1/(1 + d'_k)$ for $k \in 1, ..., K$.

In the case of categorical data the size of the intersection of the sets describing two organizations is used as the similarity between them, also known as the Jaccard index. Each organization carries a set of categories $C_i \subseteq \mathcal{U}$, where $\mathcal{U}$ contains the set of possible categories (e.g. the taxonomy). For each of the categorical parameters, we iterate over combinations in $\boldsymbol{O}$ still by indexing its elements with $1 \leq i < j \leq m$ as $\boldsymbol{j} = (j_k)_{k=1}^{K}$, where $j_k = |C_{i_k} \cap C_{j_k}| \ / \ |C_{i_k} \cup C_{j_k}|$, for $k = 1, \ldots, K$.

With the latter computation, we end up with the proportion of common elements between two organizations. Since the size $|\cdot|$ of the intersection is normalized by the size of the union in the previous equation, the categorical similarity exists by construction as $j_k \in [0, 1] \ \forall \ k$.

All similarity scores presented allow for ranking where 1 is the maximum similarity possible and 0 is the lowest. The results shown to the user display the most similar organization to another upon demand. For example, a Municipality operator can obtain a list of most similar Municipalities based on geographical and demographic parameters, and if their own Municipality has submitted plans in the past they may be included in the similarity ranking as well. In Table 2 we show the results obtained for the small city of *Aldeno* (TN). Aldeno participated in Family in Trentino, and we display its similarities using the average for geo-demographic and for past plan parameters - even though the operator has also access to the individual similarities for each parameter. The similarity can also be computed for municipalities that did not participate (yet), and may want to have a data-informed ranking of which are the most similar municipalities, given that only geo-demographic data is available. An operator can also weight the different parameters differently. For example, they may decide how much the altitude or the similarity between past actions *etc.* should impact the final result.

Table 2
Top-3 municipalities by similarity for *Aldeno* (already participating in the certification) and *Cavedago* (with no past actions submitted).

Municipality similarity w/ Aldeno (participating)	**Geo-demogr.**	**Past plans**	**Avg. score**
Ala (TN)	0.842	0.802	0.819
Ospedaletto (TN)	0.853	0.750	0.808
Campodenno (TN)	0.888	0.673	0.774
Municipality similarity w/ Cavedago (no past plan data)	**Geo-demogr.**	**Past plans**	**Avg. score**
Lavarone (TN)	0.994	N.A.	N.A.
Bleggio S.(TN)	0.989	N.A.	N.A.
Sover (TN)	0.992	N.A.	N.A.

3.2. Action Classification

The creation of family support plans foresees the description of actions that the organization commits to carry out to be family-friendly, together with the assignment of one or more categories taken from a pre-defined taxonomy, that is different in the case of companies and municipal proponents.

For classification, given the availability of enough submitted plans, we choose a supervised framework. Specifically, we fine-tune a Bidirectional Encoder Representations from Transformers (BERT) pre-trained on Italian corpora, known as BERT Base Italian XXL[2], inspired by the promising results obtained in a similar classification task in Italian [1].

The training data for the classification of actions submitted by municipalities and by companies are different, also because of differences in the underlying taxonomies. The data has been collected just over 14 years of planning activity (2008-2022). To avoid data leaks between training, validation and test sets, some preprocessing was required. First, we removed duplicates in the text descriptors and extremely similar descriptors (using an adjusted version of the tool reported in 3.3) so as to not have the tests be corrupted by the presence of identical observations both in the training and the test set. We know that some of the organizations involved would re-use actions from one previous plan to another if actions lasted multiple years, hence the presence of duplicates or extremely similar descriptors. The final dataset contains 18,102 (80 labels) observations for municipalities, and 11,483 observations (133 labels) for companies. The total number of labels does not correspond to the number of classes in the respective taxonomies because we removed the categories with less than three instances. We use the same classification framework (i.e. multiclass) with some adjustments per data. The training data from companies that actually have multiple labels are present in 34% of the observations, so the final model assigns more than one label, while Municipalities' data only have one category per observation. Given the training data, the loss of choice for training is the *Cross Entropy Loss* and its variation for

[2] https://huggingface.co/dbmdz/bert-base-italian-xxl-uncased

multi-label setting (*Binary Cross Entropy Loss*), through PyTorch'spackage [2]. In our multi-label case, where N is the batch size and the BCE is reduced to the batch mean. Given a multi-label vector $\boldsymbol{y}$, $\ell(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{N} \sum_{n=1}^{N} l_n$, where $l_n = -\, w_n \big[y_n \cdot \log(x_n) + (1 - y_n) \cdot \log(1 - x_n) \big]$ and where predictions $\boldsymbol{x}$ are the model raw output passed through a sigmoid function.

In both classification tasks, the training data is $\approx 65\%$ of the dataset, the validation set is $\approx 15\%$, and the remaining $\approx 20\%$ is the test set. They were split via stratified sampling, so that the category distribution is preserved in the three splits.

Results. The classifier performance is evaluated using $F1$ weighted by the frequency of the label classes. The computation of the metric is obtained through Sci-Kit learn [3]. In Table 4 we report the classifier performances both on the validation and the test set. As expected, multi-label classification on companies' actions is less accurate than single-label classification of municipalities' actions. However, results are rather promising considering the sheer number of labels to predict. Furthermore, this classification is not meant to replace human labeling but only to assist and speed-up the process, so we may consider presenting to final users not only one label but the list of those classified with highest confidence.

Table 3: Performance results for Italian BERT on Municipalities (single-label) and Companies (multi-label)

Metric	**Single**		**Multi**	
	val.	test	val.	test
F1 (micro)	0.754	0.729	0.701	0.688
Precision (micro)	0.755	0.731	0.743	0.746
Recall (micro)	0.753	0.727	0.663	0.638

We perform a further analysis by generating the Precision-Recall curve for the two tasks, which is displayed in Figure 1. Each point in the curve is the intersection of the precision and of the recall of the predicted labels given different thresholds for activation function. The highest F_1 obtained analyzing these curves represents the point of best trade-off between prediction and recall. According to our analysis, the best activation function threshold for the single-label trained model is relatively high: 0.94. Overall, the two curves are reasonably similar, although the best activation threshold for the multi-label model is 0.35. The fact that the single-label trained model performs best with such a high activation function threshold, far from the default 0.5, indicating that the architecture is prone to overconfidence in this setting. However, this is a known phenomenon in most BERT-based [4] and other deep learning models [5]. When this issue is diagnosed in the predictions on out of domain observations, there are usually counteracting measures that are applied in different phases of the model development. One of them is post-training temperature scaling [6], not as effective as simply changing the activation function threshold, as informed by the PR curve. On the other hand, a completely different result occurs in the multi-label setting, where labels are predicted *underconfidently*, a phenomenon that has occurred also in multi-label Bayesian Neural Networks [7]. We speculate it may be due to the specific nature of the data: since most of the multi-label observations do *not* actually have more than one label (66%), it may be that the model learned to predict the least amount of labels for observations in order to be as close as possible to the real data.

3.3. Suggestion of Changes in the Taxonomy

A third component aimed at assisting operators in using the platform is meant only for the *Agency of Social Cohesion*, to support periodic revisions of the taxonomies and perform consistency checks on the submitted plans. Specifically, given a taxonomy and the database of past submitted plans, the tool suggests categories to be removed, to be merged or to be split. The three suggestions types are computed through an algorithmic procedure that analyses past submitted plans and identifies:

1. Taxonomy categories that are used less than k times (with k being defined by the user): candidates for *deletion*
2. Taxonomy categories that are used too frequently, meaning that they may be too broad: candidates for *splitting*

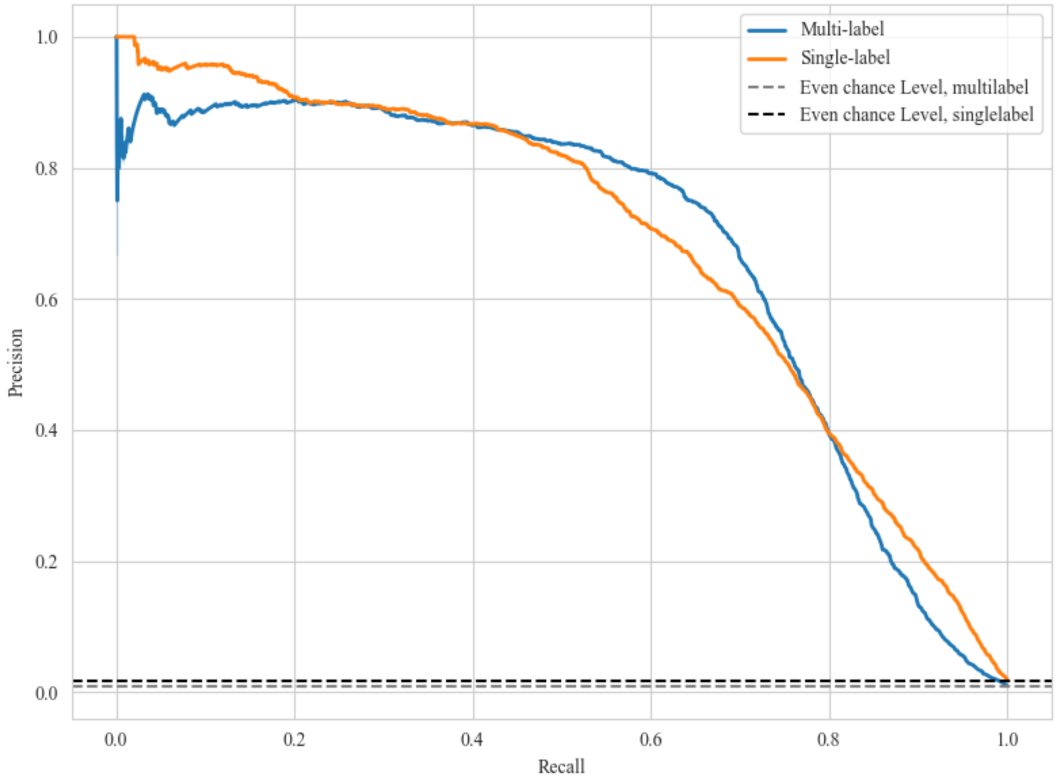

Figure 1: Precision Vs Recall, computed using predicted probabilities. The points of the PR curve vary upon different activation threshold functions, signaling the best threshold for desired trade off.

3. Taxonomy categories that are different but are used to label similar actions: candidates for *merging* or *refinement*

The first two suggestions are straightforward to compute, since they are based on a frequency analysis of the actions in submitted plans, which first identifies the most- and least-frequent actions categories and then retrieves the corresponding labels from the taxonomy. For example, if we consider the *Family in Trentino* certification, municipalities have used only once in past years the categories *Co-progettazione attivita del progetto strategico del Distretto famiglia* and *Violenza di genere: Servizi di supporto per uomini maltrattanti.* These initiatives are not popular choices and the Agency, upon inspection of the descriptive statistic, may consider removing a specific category that just covers these cases. Concerning the top-frequent categories, *Sostegno economico alle associazioni del territorio / Concessione spazi* has been used $1,146$ times in past years and the tool suggests to split it into two finer-grained categories, and just like it a handful of others have been thoroughly used. These could probably be *Sostegno economico alle associazioni del territorio* and *Concessione spazi alle associazioni del territorio.*

As regards the third type of suggestions, concerning merging or refining categories, it requires the implementation of a specific algorithm, aimed at detecting actions in the submitted plans that are similar but that have been manually labeled with different taxonomy categories. This may be due to an operators' mistake, but also to the presence in the taxonomy of classes that may be merged or revised because too similar. To compute pairwise similarity between actions descriptions we used the *fuzzywuzzy* library [3], which is based on the inverse of the Levenshtein (edit) distance between two strings, i.e. between 0 and 100. We compute a specific score named *token set ratio*: first, words are tokenized and only unique tokens are kept, then the inverse of the Levenshtein distance is computed. With this method, two strings are scored 100 if they are essentially identical texts, while greatly

[3] https://pypi.org/project/fuzzywuzzy/

dissimilar strings result in a score of ≈ 30 or lower.

More specifically, the algorithm is implemented as follows. Let the data be composed of observations with descriptors A for the actions written by operators such as descriptions, titles, etc. The data represents all participating organizations $\boldsymbol{O}$.

- For each organization O_i in set of organizations $\boldsymbol{O}$:
  1. Retrieve all elements $\boldsymbol{A}$ as descriptor-category pairs from organization O_i, where we define $\boldsymbol{A} = \big[(A_1, C_1), (A_2, C_2), ..., (A_n, C_n)\big] \quad C_k \in \mathcal{U}$
  2. As the result of symmetric function fuzzy$(\cdot)$, compute vector $\boldsymbol{v}$ containing pairwise *fuzzywuzzy* score $\boldsymbol{v} = \big\{(\text{fuzzy}(A_p, A_q), C_p, C_q) \mid 1 \leq p < q \leq K\big\}$ over all possible unique sorted combinations of action descriptors in $\boldsymbol{A} = (A_k)_{k=1}^{K}$.
  3. Filter the tuples in $\boldsymbol{v}$ based on *fuzzywuzzy* score with an arbitrary threshold t denoted as $\boldsymbol{v}'$
  4. Return $\boldsymbol{v}'$
- After iterating for all organizations, return set $\boldsymbol{V}$, composed by all vectors with triples $\boldsymbol{V} = \bigcup \boldsymbol{v}'_i$.

The frequencies of matches inserted in V are be counted for aggregative measures indicating the overall usage of categories and actions descriptors for all participating organizations. On a computational expense note, the algorithm complexity is highly dependent on the length of the strings and the distance used in item 2, even if it is lessened by lightly preprocessing on the text (stripping strings from punctuation and Italian *stop words*), and processing using *fuzzywuzzy*'s set method (removes duplicate elements from the string). However, it is using as much time as the average case scenario [8], even though it is more efficient than its "sibling" method Sequence matching [9]. We obtained similar results using smaller strings: titles, instead of descriptions, are shorter descriptors, and while some information is lost, they are usually still indicative enough to indicate miscategorisation. In the step 3, our threshold for satisfactory similarity is $t = 75$, in which most of the elements of the descriptor match other's, besides a few changes. We compare actions submitted over the years by the same organisations because it is more likely to find similar descriptions compared to different organisations. While a lower threshold contributes to avoid redundant or undesirable matches, one can also set an upper boundary in case exactly identical or too similar descriptors are not of interest.

The use of this metric has proven to yield interesting results. For instance, actions with high similarity have been found almost 50 times in the plans submitted by municipalities as being labeled with two different labels, either *Servizi doposcuola e servizi estivi* or *Centri di aggregazione per giovani.* This suggests that the two categories in the taxonomy should be revised by domain experts and possibly redefined. This revision of the data is useful also to improve the classifier performance, since similar descriptions bearing different labels would probably introduce noise when training the classification model.

4. Conclusions and Future work

In this work we presented a series of tools for data-informed decisions in the public administration domain. In particular, we detailed three components that are meant to support the process undertaken by municipalities and companies to obtain a family-friendly certification. We implemented NLP tools for text classification, similarity computation and taxonomy analysis that to our knowledge were never applied to this domain before. While their first implementation is completed, they still require further qualitative testing and improvements in terms of robustness.

The similarity tool described in Section 3.1 is not only designed for the user to visualize the most similar organizations, but it is also meant to be integrated into another tool – LLM-based dialogue system for the preparation of plans. Given an organization of interest, the system shall retrieve the plans submitted by the most similar organizations, as a basis to inform text generation with more specific context and documentation. The tool to compute similarity indices between organizations will be extended in the future to include new sources of information, especially related to services available in different municipalities. Its statistical expressive power may also be refined by implementing more robust kernel similarity measures, as the (inverse) of an euclidean distance is a dissimilarity measure [10] but does not correspond to some inner product in the feature space. The deep learning classifier reported in Section 3.2 is publicly available for further training and reuse.[4] Given its flexibility, it can be easily retrained to adapt to changes in the taxonomy or to integrate updates of the database of submitted plans. It may also be used in other projects that share the same data setting: Italian-language descriptors, action-based datapoints, and a large taxonomy. Its overconfidence shall be tackled using adaptive measures to prevent it while training, such as label smoothing, loss regularization [11], and tempered loss [12].

Ultimately, the quality of the tools will be assessed by stakeholders from *Agency for Social Cohesion* and municipalities during five stakeholders' meetings before the end of the project (Fall 2025).

Acknowledgments

This work is part of the project "AIxPA – Artificial Intelligence for Public Administration" , funded within the Flagship Project PNC-A.1.3 – Digitalization of the Public Administration of the Autonomous Province of Trento, CUP: C49G2200102000.

Declaration on Generative AI

During the preparation of this work, the author(s) used Chat-GPT in order to: format tables and tidy command syntax in LaTeX. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] M. Rovera, A. P. Aprosio, F. Greco, M. Lucchese, S. Tonelli, A. Antetomaso, Italian legislative text classification for gazzetta ufficiale, in: Proceedings of the Natural Legal Language Processing Workshop 2023, 2023, pp. 44–50.

[2] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems 32, Curran Associates, Inc., 2019, pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[3] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Müller, J. Nothman, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in python (2012). URL: https://arxiv.org/abs/1201.0490. doi:10.48550/ARXIV.1201.0490.

[4] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On calibration of modern neural networks (2017). arXiv:1706.04599.

[5] H. Wei, R. Xie, H. Cheng, L. Feng, B. An, Y. Li, Mitigating neural network overconfidence with logit normalization, in: K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, S. Sabato (Eds.), Proceedings of the 39th International Conference on Machine Learning, volume 162 of *Proceedings of Machine Learning Research*, PMLR, 2022, pp. 23631–23644.

[6] A. S. Mozafari, H. S. Gomes, W. Leão, S. Janny, C. Gagné, Attended temperature scaling: A practical approach for calibrating deep neural networks (2018). arXiv:1810.11586.

[4] https://github.com/FluveFV/multilabel-aixpa

[7] F. Rewicki, J. Gawlikowski, Estimating uncertainty of deep learning multi-label classifications using laplace approximation, in: IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium, IEEE, 2022, pp. 1560–1563.

[8] G. Navarro, A guided tour to approximate string matching, ACM Computing Surveys 33 (2001) 31–88. doi:10.1145/375360.375365.

[9] G. A. Rao, G. Srinivas, K. Rao, P. P. Reddy, Characteristic mining of mathematical formulas from document - a comparative study on sequence matcher and levenshtein distance procedure, International Journal of Computer Sciences and Engineering 6 (2018) 400–404. URL: http://dx.doi.org/10.26438/ijcse/v6i4.400404. doi:10.26438/ijcse/v6i4.400404.

[10] C. Scheidt, J. Caers, Representing spatial uncertainty using distances and kernels, Math. Geosci. 41 (2009) 397–419.

[11] D.-B. Wang, L. Feng, M.-L. Zhang, Rethinking calibration of deep neural networks: Do not be afraid of overconfidence, in: M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, J. W. Vaughan (Eds.), Advances in Neural Information Processing Systems, volume 34, Curran Associates, Inc., 2021, pp. 11809–11820. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/61f3a6dbc9120ea78ef75544826c814e-Paper.pdf.

[12] F. Wang, S. Mizrachi, M. Beladev, G. Nadav, G. Amsalem, K. L. Assaraf, H. H. Boker, MuMIC – multimodal embedding for multi-label image classification with tempered sigmoid (2022). arXiv:2211.05232.