Computer Science and Information Systems 23(2):917–946 https://doi.org/10.2298/CSIS251029022A

A framework for the automated thematic annotation of Open Government Data

Abdul Aziz[1], Mohsan Ali[2], Dagoberto José Herrera-Murillo[1], Maria Ioanna Maratsi[2], Francisco J. Lopez-Pellicer[1], and Javier Nogueras-Iso[1]

[1] Aragon Institute of Engineering Research (I3A), Universidad de Zaragoza, Zaragoza, Spain
{abdul.aziz, dherrera, fjlopez}@unizar.es
jnog@unizar.es (corresponding author)
[2] Department of Information and Communication Systems Engineering, University of the Aegean, Samos, Greece
{mohsan, ioanna.m}@aegean.gr

Abstract. Governmental policies for transparency and reuse of public sector information have encouraged the launch of open government data portals around the world. Many of these portals are based on pyramidal structures: national open data portals are aggregators of the contents harvested from open data portals maintained by governments in charge of administrative areas with a narrower scope. Taking into account this hierarchical organization, these open data portals lack consistent and scalable mechanisms for thematic annotation, limiting dataset discoverability. This work proposes a framework for the automated thematic classification of open government data. The framework integrates (i) thematic annotation quality assessment, (ii) supervised machine learning models trained on annotated metadata corpora, and (iii) embedding-based semantic similarity methods for theme assignment in the absence of reliable annotations. The framework is evaluated using 29,793 datasets from *data.europa.eu*, the European open data portal. Experimental results show that supervised models achieve high classification performance, with Support Vector Machines reaching an accuracy of 93.65%, while unsupervised embedding-based approaches achieve substantial semantic agreement with portal-assigned themes (74.56%) using transformer-based representations. These results demonstrate that the proposed framework enables scalable, consistent, and interoperable thematic annotation, offering both theoretical contributions to automated metadata enrichment and practical value for integration into large-scale open data portal infrastructures.

Keywords: Data Annotation, Open Government Data, Open Data Portals, Automated Annotation, Thematic Annotation

1. Introduction

In the era of digital governance, governments worldwide are increasingly adopting open data initiatives to facilitate the access to open government data (OGD) [25,17,33]. OGD, as a subset of the broader concept of open data (data in open format to be shared, used and reused for any purpose), covers a wide range of topics - from budget records to environmental monitoring among other examples. OGD have the potential to enhance transparency, foster citizen participation, support evidence-based policymaking, and drive innovation through the use and reuse of data [34,32,37]. The increased deployment of open

data catalogues and portals has enabled the distribution and straightforward retrieval of substantial quantities of open data in the information-driven society and has leveraged the growth of the open data movement [53,4,7,26].

However, the challenge lies in making sense of this enormous resource. OGD portals are usually organized on the basis of pyramidal structures: national open data portals are aggregators of the contents harvested from open data portals maintained by governments in charge of administrative areas with a narrower scope. This means that open data catalogues must usually integrate heterogeneous metadata records describing datasets that have been harvested from different catalogues with a more local scope. Therefore, often open data catalogues struggle with findability due to limited and inconsistent metadata [7,3].

Taking into account these scenarios, where OGD portals at national or cross-national level must integrate the metadata contents from different sources, it is essential to provide users with classification tools in order to easily locate the information of interest [10]. Information classification schemes generally fall into two main types: exact and ambiguous [38]. Exact organizational schemes precisely divide information into clear and mutually exclusive sections, using methods such as alphabetical, chronological, or geographical sorting. Ambiguous organizational schemes, on the other hand, delimit information into categories that resist precise definition, often due to linguistic ambiguities and human subjectivity. This classification encompasses thematic organization schemas along with other variants such as task-oriented and metaphor-oriented schemas. Thematic organization, in particular, prioritizes the grouping of related items, promoting associative learning, and facilitating adaptive information retrieval strategies.

This work is focused on the study of the thematic annotation of datasets included in OGD catalogues. Rich metadata is one of the core tools to fulfill the FAIR principles [62] and improve the findability, accessibility, interoperability, and reuse of digital assets. To address the findability aspect, one of the almost mandatory recommendations is to include keywords and themes within metadata. This is typically checked by metadata quality evaluation methodologies [30,42,61]. Although national or even cross-national catalogues like *data.europa.eu*, the official European Data Portal, combines metadata compliant with different metadata vocabularies, most of them are derived from DCAT [28,11]. DCAT is the W3C's Data Catalog vocabulary for describing open data [60]. The advantage of using DCAT derived vocabularies is that the thematic annotation of datasets is encouraged thanks to the inclusion of a specific property called *dcat:theme*. Within the metadata, themes provide a higher degree of semantic structure that goes beyond individual keywords and descriptions. Users may conveniently find important information by classifying the datasets according to their applicable themes, independently of the use of exact terminology.

Nevertheless, just the use of metadata vocabularies including a property for thematic annotation is not enough. A missing or wrong thematic annotation hinders the findability. Therefore, we must ensure that the content of this property is accurate. Recognizing the potential of thematic annotation, several researchers have investigated several techniques for annotation, which involves giving significant labels to data. Manual annotation is precise and thorough but is limited in its capacity to scale up [6,64,47]. Automatic annotation exploits machine learning and natural language processing (NLP) to automatically classify information thematically [19]. Using automated annotation may have several advan-

tages. It has the ability to significantly reduce the amount of human work needed, enabling the rapid processing of large datasets. Moreover, it may promote consistency and establish criteria in the annotation process, hence improving the findability and availability of the annotated resources [51].

While there has been substantial progress in publishing OGD, many OGD portals still lack a consistent and scalable mechanism for the thematic annotation of datasets coming from heterogeneous sources, making it difficult for users to discover relevant datasets efficiently. Existing approaches often rely on manual categorization and are inconsistent across portals. This need for an automated mechanism establishes a research niche, which is covered in this work with the design of a framework that encompasses a structured and systematic set of concepts, methods, and software components to guide the process of thematic annotation of OGD. The design of this framework aims to address three main research questions:

1. What is the current state of thematic annotation in open government datasets, and how accurately do these annotations align with the content of the datasets? To answer this question, we have proposed an implementation of the method proposed by Nogueras-Iso et al. [42] for evaluating the thematic classification correctness.
2. Assuming that our collection of datasets (corpus) is properly annotated with themes, to what extent can new datasets be automatically classified? To answer this question, we have tested different machine learning algorithms and preprocessing strategies on an annotated metadata corpus.
3. In the case of not having an annotated corpus, or in the case our corpus is not properly annotated, how can relevant themes be assigned to a dataset from free text metadata? To answer this question, we have tested different strategies based on word-embeddings and sentence-embeddings of metadata to identify the closest theme from a predefined list of themes based on their definitions.

The rest of this paper is structured as follows. Section 2 provides a literature review about the thematic annotation of OGD. Then Section 3 presents our proposed framework for the automated thematic annotation of OGD, which includes the evaluation of the thematic classification correctness in the case of having an existent annotated corpus. Section 4 presents the results after applying the proposed methodology to a corpus of metadata records from the European Data portal (*data.europa.eu*), which are discussed in Section 5. Finally, this paper concludes with a summary of the contributions and some ideas for future work.

2. Related Research

Metadata is a critical component in numerous facets of data management, encompassing the integration, transmission, and transformation of data, among others [43]. As highlighted by the frameworks for assessing the quality of open data portals [30,42,61,49], missing or incomplete metadata hinders the findability of data. A wrongly assigned theme that does not accurately represent the content of a dataset reduces its discoverability and search recall for stakeholders, thereby motivating the need for efficient and accurate thematic annotation and negatively affecting the usability and usefulness of open government

data portals [5]. User-centered studies of open data portals have similarly reported difficulties in dataset discovery and navigation, often linked to metadata quality and categorization practices [41], reinforcing the need for systematic thematic annotation approaches.

Given the importance of providing correct metadata without the burden of accomplishing this task manually, different strategies have been suggested and developed over the years with the aim of achieving a fully or partially automated thematic annotation of resources. Although each strategy emphasizes a unique set of conceptual areas of knowledge and experience, artificial intelligence techniques like machine learning are acquiring an increasing role of portal curators [54,44,1].

As Semantic Web technologies are widely used as a mechanism to publish and reuse open data [15] and metadata is the core of the Semantic Web [58], many research works on automatic annotation are close related to the use of these technologies. For instance, Pavia et al. [45] applied ensemble methods to classify Web-scale datasets through their metadata using a hybrid Recurrent Neural Network composed of LSTM and Bi-directional LSTM units and Naïve Bayes models at a second phase. In a more specific context and regarding bibliographic data, Carducci et al. [9] worked on text categorization for automatic metadata annotation in order to annotate records, separating between philosophical documents and other disciplines. To facilitate this binary classification purpose, they employed NLP and other ensemble learning techniques, integrating domain knowledge and information gained through semantic networks (BabelNet) to decide whether a given document (e.g., thesis) is within the philosophical domain or not. The annotated data is then used to train the chosen supervised learning algorithms and automatically classify the metadata according to the thematic subject of the examined record. Likewise, Verberne et al. [59] investigated the processing and classification of electoral manifestos. After optimizing different parameters including passage segmentation, OCR, or formatting, the results showed that the classifier matches human experts in accuracy and recall.

There are also recent studies focused on OGD portals highlighting the role of automated keyword extraction in enhancing thematic organization and improving findability in open data portals. For instance, Ahmed et al. [2] proposed BRYT, an automated keyword extraction tool that merges and select the most prominent keywords obtained by different techniques based on the statistical distribution of words in the metadata and Large Language Models (LLM). Similarly, Kliimask and Nikiforova [29] introduced TAGIFY, a language model-powered tagging interface aimed at improving data discoverability through enriched metadata in OGD portals. Freire et al. [18] analyzed the use of LLMs in diverse data integration and data discovery tasks, synthesizing recent advancements in this rapidly growing domain. Moreover, Zhang et al. [65] introduced AutoDDG, a framework that is explicitly designed for the automated generation of dataset descriptions for tabular data. AutoDDG utilizes a data-driven methodology to provide a concise summary of dataset content. It employs LLMs to enrich these summaries with semantic information and to generate comprehensible descriptions. These approaches underscore how NLP and semantic tagging can mitigate linguistic ambiguities inherent in ambiguous organizational schemes, thereby supporting more effective associative learning and adaptive retrieval strategies. Huseynov et al. [22] also emphasized the power of NLP to propose a recommender system for datasets. Using the Word2Vec word-embedding technique to encode the free text content of different metadata properties in a vector space, their system provides the users with the possibility of selecting an input dataset and discovering the

recommended datasets with a closer embedding in the vector space. Somehow connected to recommender systems, Bogdanovich et al [8] proposed a method based on Formal Concept Analysis to create a lattice of keywords using as input source the tags for describing datasets under the same thematic category but hosted in different open data portals. This lattice of keywords (concepts) allows cross-portal search of related datasets.

Several attempts for improved annotation services using semantic approaches have also been made in specific data domains such as the biomedical domain. Sasse et al. [52] conducted a literature review on existing semantic metadata annotation services and identified their software requirements in accordance with the FAIR principles: availability as open code; compatibility with common data formats; use of FAIR terminologies; possibility of terminology search; suggestion of annotations; availability of interfaces to external terminologies; and extension of terminologies. Although they concluded that there are not metadata annotation tools that meet all the requirements, this study highlights the importance of annotation tools and the availability of functionalities for suggesting annotations. In a more specific context about the psychiatric and psychological domain, Hudon et al. [20] analyzed the literature on the potential of machine learning to assist in the thematic annotation and classification of text in a psycho-therapeutic context. Their findings demonstrated that, although the existing literature on this specific topic is limited, some techniques such as Support Vector Machine classifiers achieved sufficient accuracy in the performed text classifications, and that this type of classifier is consistently used for classification in the context of medical or clinical text data [21].

Automatic annotation has also been attempted for environmental science metadata. Tuarob et al. [56] aimed to alleviate the problem of environmental metadata harvesting from various and disparate sources with varying levels of metadata quality and curation. They gathered datasets from 4 different archives, selecting for each of them a subset of 1000 annotated documents, and the textual content and attributes of the documents were pre-processed (removal of stop-words, stemming etc.) to obtain a *tf-idf* (term frequency - inverse document frequency) representation. In order to rank automatically candidate themes for the dataset, they used different similarity measures based on cosine similarity and Latent Dirichlet Allocation. Focusing on the processing of images, Ellen et al. [14] targeted plankton image classification using context metadata (such as perimeter, symmetry, temporal and geographic information, etc.) in order to improve the performance of feature-based classifiers. They demonstrated that the inclusion of context metadata might be of substantial gain for classification accuracy in deep learning models, mainly Convolutional Neural Networks. Likewise, Peng et al. [46] proposed a unique biological data classification feature selection method to enhance feature categorization. The technique uses filter and wrapper approaches: it pre-selects feature subsets to improve search efficiency and utilizes ROC curves to assess feature and subset performance. Furthermore, on the viticulture domain, Mylonas et al. [39] proposed a platform for data annotation that includes a thesauri manager for the obtainment of Linked Data Vocabularies. These vocabularies are used in the platform for both manual and automatic annotation based on NLP techniques and supervised learning models such as *k-nearest* neighbors and linear and random forest regression.

When domain-specific research is being carried out, the specificities and domain-sensitive requirements need to be taken into account, to prevent or be aware of in-advance algorithmic biases and limitations. Wu et al. [63] presented the status for automated meta-

data annotation in the cultural heritage domain and discussed the potential of machine learning applications supporting the curating processes of digital artifacts. They provided a summary of recommendations to improve these aspects of automated metadata annotation by leveraging already existing text and images of high quality, utilizing inference of meaning for classification from simple object recognition to tackle metaphoric and symbolic representations in the digital realm, and providing quality indicators on the results to tackle non-uniform and non-consistent automated indexing. Similarly, Ibáñez et al. [23] provided a quantitative analysis of Linked Data in accessible government datasets throughout Europe. They examined the popularity of RDF as a publication format, the accuracy of connected datasets, and the prevalence of established terminologies. Furthermore, the negative effect of poor metadata description on the discoverability of digital cultural heritage artifacts was also addressed by Kaldeli et al. [27] who proposed CrowdHeritage, an ecosystem supportive of end-to-end improvement of metadata utilizing crowdsourcing, machine and human intelligence, semantic, and aggregation techniques.

The framework for thematic annotation proposed in this work integrates the existing knowledge in the state of the art of this field. First, the initial assessment of thematic classification correctness adapts the methodology proposed by Nogueras-Iso et al. [42] to establish quality controls on dataset themes. Second, the supervised classification techniques applied for new datasets in case of having a previously annotated corpus are similar to other works in the literature [45,21]. In addition, the needs for preprocessing and feature representation are similar to other works using free text metadata as input [56]. Third, the unsupervised classification techniques applied in our framework also share some similarities with respect to the works of Ahmed et al. [2], Kliimask and Nikiforova [29] and Huseynov et al. [22] as they also exploit the benefits of using word embeddings and language models. Our proposal compiles all these alternatives within a unified framework, which allows the comparison of the suggested thematic annotations for new datasets in two different scenarios: the existence of a properly annotated corpus; or the unavailability of a properly annotated corpus.

3. Methodology

This section outlines our proposed methodology of the thematic annotation of OGD. Figure 1 shows the general workflow envisioned in this framework. In the case of counting on an annotated corpus, we first need to evaluate the thematic classification correctness before building a machine learning model for the classification of datasets. In contrast, if there is not an available annotated corpus or its classification correctness is not acceptable, we opt for predicting the closest theme measuring the similarity between the word/sentence embeddings of datasets and themes.

3.1. Evaluation of thematic classification correctness

As we need an annotated corpus for the ulterior development of automatic classification models, it is necessary to evaluate first the thematic classification correctness of the corpus. For this evaluation we propose to follow the method proposed by Nogueras-Iso et al. [42]. This method is a customization of the original method proposed by Ureña-Cámara et al. [57], which adapts ISO 19157 standard for geographic information quality

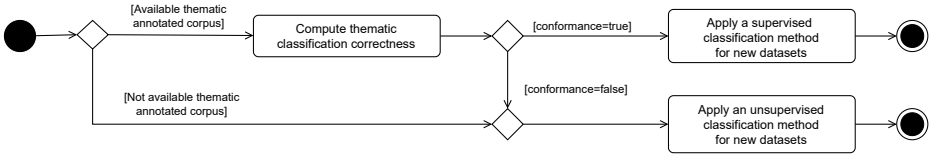


Fig. 1. Workflow for the automated thematic annotation of OGD

.

for the assessment of geographic metadata compliant with the ISO 19115 metadata standard. Nogueras-Iso et al. describe how to perform the assessment of open data metadata compliant with DCAT-based metadata models serialized in RDF. They propose a series of quality controls on six quality categories: completeness, logical consistency, temporal quality, thematic accuracy, positional correctness, and quality of free text. In particular, the thematic accuracy category includes a quality element focused on thematic classification correctness, i.e., the correctness of the thematic keywords and categories included in the metadata with respect to a universe of discourse.

The assessment of the thematic correctness must be made using a sample-based inspection and a *Limiting Quality* index, which determines the sample size (n) according to the corpus size and the maximum number of errors (Ac) that can be accepted to assure a statically equivalent percentage of errors (*Acceptance Quality Limit* or *AQL*) if the full corpus were evaluated. Therefore, the computation of the thematic classification correctness requires the compilation of two associated results: a quantitative result and a conformance result. The quantitative result consists in obtaining a numerical value for the ISO 19157 D.63 measure, which is defined as the number of incorrectly classified records. The conformance result verifies whether the number of errors in the quantitative result surpasses or not the acceptable number of errors (Ac) for the considered sample size.

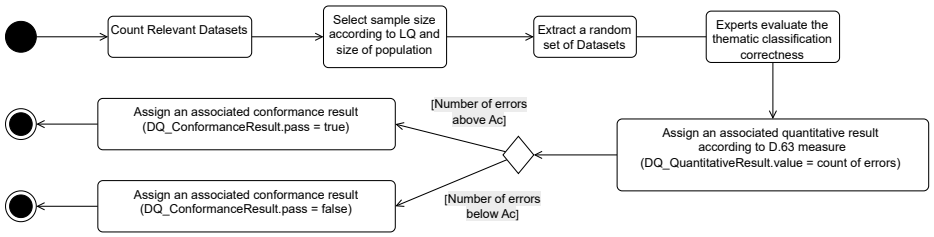


Fig. 2. The workflow for reporting the thematic classification correctness.

Figure 2 shows the workflow that must be followed to compute the thematic classification correctness. In general, the workflow of the assessment starts by considering the whole corpus of datasets of our case, including all the valid and relevant samples. Then, a selected sample size proportionate to the initial population is included to undergo the pre-assessment process, and a random sample set is chosen using a random number gener-

ator. Afterwards, the process of manual assessment of instances from the selected sample is initiated. In order to implement this assessment, we decided that the random sample had to be evaluated by different experts and that this evaluation implied the inspection of the resources associated with the datasets. Then, the experts should assign to them between one and three related themes according to the perceived content of the dataset, its title, its description, and the associated keywords. Then, a consensus should be reached by the experts to consider correct a theme classification if at least one of the assigned themes by the experts corresponded to the initial theme assigned to the dataset. The cases where the initially assigned theme of the dataset does not correspond to the themes assigned by the experts should be annotated as errors. Finally, the associated quantitative and conformance results are assigned.

3.2. Learning to automatically classify based on annotated corpus

Assuming that we count on an annotated corpus of datasets where each dataset has been properly annotated with themes, this component of our annotation framework is focused on building models for the automatic thematic annotation of datasets. For this purpose, we have tested different machine learning algorithms that are typically applied for automatic classification problems in supervised scenarios. Figure 3 shows the workflow followed to build a model for the thematic annotation of OGD thematic annotation. The proposed steps in this workflow are the selection of metadata properties, the normalization of the input text to extract terms, the transformation of the terms into an appropriate feature representation, and the generation of the classification models.

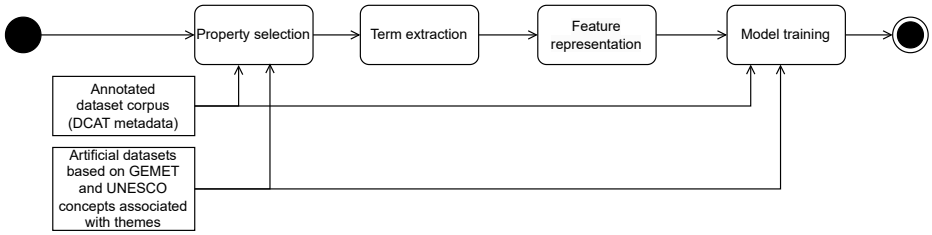


Fig. 3. Proposed method for automated classification of open datasets based on an annotated corpus.

Property selection In this framework, we assume that metadata is compliant with a metadata vocabulary derived from DCAT. As shown in Figure 4, this type of vocabularies include free-text properties for describing a dataset in terms of a title (*dct:title*), a general description (*dct:description*) and several keywords (*dcat:keyword*). In addition, datasets have also an associated theme thanks to the *dcat:theme* property, whose range is a concept from a well-known Knowledge Organization System (KOS) expressed in SKOS format [36]. Therefore, we decided to use the combination of the text provided in *dct:title*, *dct:description* and *dcat:keyword* as input text for the classification. With respect

to the themes or categories to be assigned after the classification process, we assumed in the experiments that the list of themes belonged to the KOS proposed by the European data portal [48], but this is interchangeable with any other KOS if a different corpus of metadata must be classified.

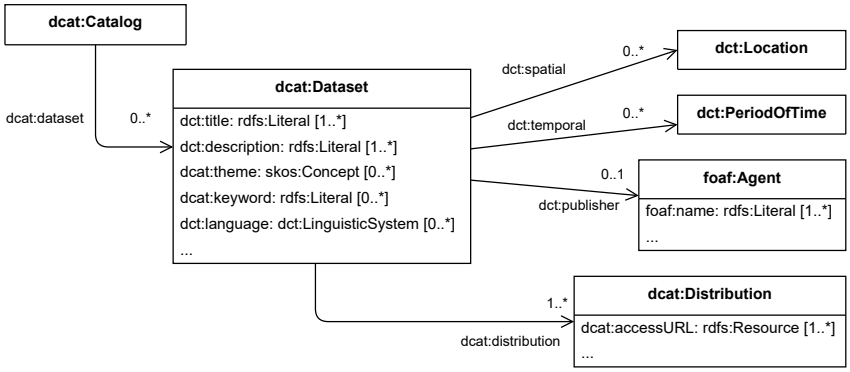


Fig. 4. An excerpt of DCAT-AP metadata model highlighting the free-text elements for describing datasets and its thematic classification (*dcat:theme*). DCAT-AP [16] is one of the main application profiles derived from DCAT for the description of public sector information.

In addition, we also considered the possibility of generating some artificial datasets for each theme to reinforce the classification models to be built. For this research study, we selected GEMET[3] and UNESCO[4] thesauri due to their thematic breadth and multilingual coverage. These thesauri have been widely used for cataloging purposes over the years to facilitate a harmonized thematic classification of datasets and reduce the gap between the vocabulary of data users and data publishers [13,35]. Using the GEMET and the UNESCO thesauri, we aligned the European data themes with the main themes in GEMET and UNESCO (a theme is defined as a micro-thesaurus in UNESCO). Using this alignment, we extracted the preferred labels of the concepts associated with each theme in GEMET and UNESCO thesauri. Dividing the list of associated concepts for each theme in groups of a fixed number of concepts, we converted each group of concepts into an artificial dataset classified with a European data theme and described with the preferred labels of these concepts.

Term extraction The next step in the workflow is the tokenization of the input text describing each dataset and the extraction of terms. For the transformation of tokens into final terms, we have considered a mandatory *basic* level of normalization, and two optional processes of normalization called *translation* and *tailored* normalization.

The mandatory *basic* normalization level incorporates the following processes: stop word removal (i.e., removing common words like ‘a’, ‘an’, ‘the’, etc.), special character

[3] https://www.eionet.europa.eu/gemet/en/themes/
[4] https://vocabularies.unesco.org/browser/thesaurus/en/groups

removal (i.e., removing characters like '$', '%', '&', etc.), link removal (i.e., removing hyperlinks), lowercasing (i.e., converting all text to lowercase) and stemming (i.e., reducing words to their root form).

In addition, we observed that although metadata from OGD catalogues can be downloaded in RDF format and the language of metadata properties can be restricted to a common language such as English, the free-text content frequently appears in other languages. To address this issue, we explored the use of a *translation* normalization approach that employs an API to detect the most likely source language in the free text values and translate them into English, thereby improving consistency.

Last, we also considered a *tailored* normalization to remove noise in the free text derived frequently from spelling mistakes and the use of non-common English words such as acronyms, the names of data provider organizations or other technical terms which only make sense within the context of the data provider organization. For this purpose, there are resources like PyEnchant,[5] which provides access to a dictionary of the English dialects spoken in different regions of the world such as American English, British English, or Australian English and can be used to discard terms not contained in this dictionary.

Feature representation Feature representation is an essential step in our workflow since our objective is to convert unprocessed text input into a vector representation acceptable for our machine learning models. Here, we will explore key features commonly used in text processing, specifically unigrams, bigrams, and trigrams (also called *n-grams*) for all our experiments. Unigrams are typically bag-of-words vector representations where each word is a distinct dimension. Bigrams are vector representations where each dimension is a biword found in the input text. Trigrams are vector representations where each dimension is a distinct trigram. Figure 5 illustrates an example of the unigrams, bigrams, and trigrams that can be generated from an input text.

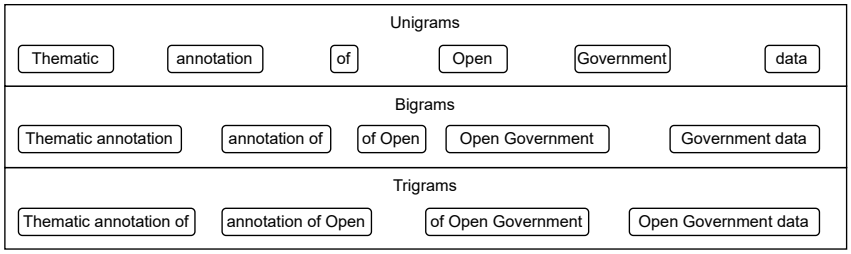


Fig. 5. An example of unigrams, bigrams and trigrams for the sentence "Thematic Annotation of Open Government Data".

N-grams do not require any typical type of calculation performed using equations or formulas. Basically, *n-gram* features are constructed by calculating the word sequences in the corpus. These number of N-grams affect the unique number of features fed into the final model training. For example, unigram features would be fewer in number than if we

[5] `https://pypi.org/project/pyenchant/`

combined unigram with bigram features. As the number of *n-grams* increases, the vector length (sparsity) increases, which increases the space and time complexity of the model training. On the other hand, there is not a standard way to decide what value of *n* for the *n-grams* will work optimally.

Model training The critical step of the workflow is the training of models where the system learns from the labelled cases (datasets annotated with themes). The models recognize patterns and properties that divide various classes, and this allows to generalize the problem and classify new, unlabeled datasets.

In particular, we have used the One-vs-Rest (OvR) classifier with three machine learning techniques: Logistic Regression (LR), Multinomial Naïve Bayes (MNB), and Support Vector Machines (SVM). The OvR classifier is used in machine learning for situations involving multiple class classifications, because it partitions multi-class problems with more than two classes into a series of binary classification tasks. Each class has its own binary classifier that has been trained to distinguish it from the others. However, data class imbalances may hurt its performance on under-represented groups [55].

Our underlying problem is to classify the open datasets not just into multi-class but also into multiple multi-class themes. For instance, a dataset in the European data portal could be classified into more than one theme of the 13 proposed themes. The OvR classifier can help us to train MNB, SVM, and LR for multiple, multi-class classification.

3.3. Predicting the closest theme of a dataset based on word/sentence embeddings

The objective of this component of the thematic framework is to predict the correct theme when an annotated corpus is unavailable or the datasets in this corpus are not properly annotated. Figure 6 shows the proposed method for the prediction of the closest theme of a dataset based on the similarity between the word/sentence embeddings representing a dataset and its potential associated themes.

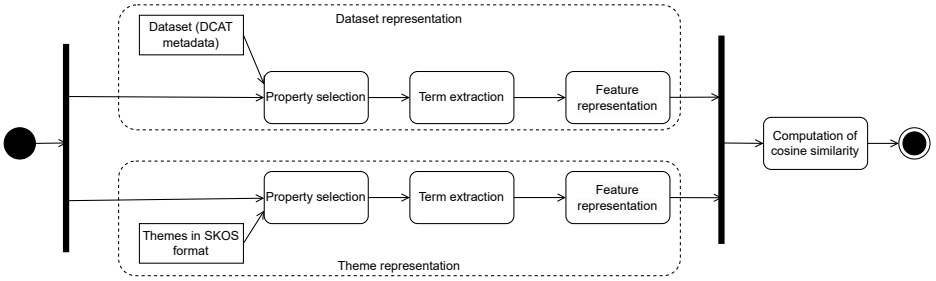


Fig. 6. Proposed method for the prediction of the closest theme of a dataset based on the similarity of word/sentence embeddings

We propose the use of word and sentence embeddings for representing datasets and themes instead of a bag-of-words representation because this allows us to represent similar input texts as close points in a vector space with a number of dimensions much lower

than the number of dimensions needed in vector spaces generated when using bag-of-words representations. Furthermore, the training data and neural networks used to generate these embeddings allow us to find similarities between two texts even in the case of not having any lexical matching between the compared texts.

Next subsections explain the process proposed for the property selection, term extraction, feature representation of both datasets and themes. In addition, we describe how we have computed the cosine similarity between the dataset and theme embeddings.

Property selection In the case of datasets, the property selection is similar to the one proposed for an annotated corpus of datasets in section 3.2. We assume that metadata is compliant with a DCAT vocabulary and we select the content of *dct:title*, *dct:description* and *dcat:keyword*. In addition, we considered two possibilities for generating the input text of a dataset: the concatenation of the three properties, and just the concatenation of title and description. We considered the second possibility because some sentence embeddings may not work properly if we create sentences including disconnected keywords.

In the case of themes, we assume that the list of themes is provided in SKOS format and that each theme is represented with an SKOS concept having an associated preferred label (*skos:prefLabel*) and a definition (*skos:definition*). Therefore, the input text to be processed for each theme is the concatenation of its preferred label and its definition in English.

Term extraction The next step in the proposed method is the extraction of tokens from the input texts for datasets and themes and the generation of terms. In this case we considered a basic normalization consisting in the removal of special characters and the transformation of text to lower case. It must be taken into account that the word/sentence embedding representation avoids implicitly the appearance of non-common English words. In addition, in some cases we also considered the removal of stop words.

Feature representation For the representation of datasets and themes, we considered different possibilities of embeddings:

- Sum of GloVe word embeddings: The terms extracted from the input text of each dataset and theme are converted into a word embedding according to the Global Vectors for Word Representation (GloVe)[6] using vectors of 200 dimensions. To represent the complete input text, this alternative computes the sum of the word embeddings.
- Average of GloVe word embeddings: This alternative is similar to the previous one, but in this case the complete input text is represented with the average of the word embeddings.
- BERT sentence embeddings: This alternative transforms the input text into a vector representation of the sentence by applying the pretrained Bidirectional Encoder Representations from Transformers (BERT) [12].
- HuggingFace sentence embeddings: This alternative transforms the input text into a sentence embedding thanks to HuggingFace representation [31,40].

[6] https://nlp.stanford.edu/projects/glove/

Cosine similarity To find the closest themes that can be associated with a dataset, we propose the use of the cosine similarity distance, which is typically applied to compute the ranking of results in information retrieval systems using a vector space model for representing documents and queries. Equation 1 shows the customization of this cosine distance to our context: the similarity between a theme T and a dataset D is equivalent to the cosine of the angle formed by the vectors $\vec{T}$ and $\vec{D}$ corresponding to their word/sentence embeddings. The similarity is therefore a real value between 0 (least similarity) and 1 (most similarity), which is computed dividing the scalar product of the embedding vectors by the product of their norms.

$$Similarity(T, D) = Cosine(\vec{T}, \vec{D}) = \frac{\vec{T} \cdot \vec{D}}{\|\vec{T}\|\|\vec{D}\|} \tag{1}$$

As the similarity is computed for all datasets that require annotation and all the candidate themes, the output of this step is a matrix where each row represents a dataset and the similarity of each theme is provided in the columns. This way we can generate a rank of associated themes for each dataset, and select, for instance, the top 3 themes.

4. Experiments and results

This section describes the applicability of the thematic annotation framework to a corpus of metadata records downloaded from *data.europa.eu*, the official portal for European OGD. The implementation of the thematic framework (Python programs and notebooks), together with the data and the associated results, are available in Zenodo.[7]

4.1. Corpus description

The metadata used in our experiments came from *data.europa.eu*. This portal serves as a centralized access point to open data published by both European Union institutions and member states. The metadata describing the datasets is compliant with the DCAT-AP vocabulary [16] and can be queried through an SPARQL end-point.[8] In July 2022 we developed a harvester program to download a corpus of 29,793 metadata records in RDF format containing title (*dct:title*), description (*dct:description*), theme (*dcat:theme*) and keyword (*dcat:keyword*) properties. One of the constraints applied to filter the corpus was to have metadata records with at least one associated theme from the list of themes proposed by the European data portal. We also restricted the download to the metadata records declaring the use of English as language, and having at least one title and one description in English.

Figure 7 shows the distribution of the datasets in the corpus among the thirteen thematic categories of the European Data Portal: 'Agriculture, fisheries, forestry and food' (AGRI), 'Economy and finance' (ECON); 'Education, culture and sport' (EDUC), 'Energy' (ENER), 'Environment' (ENVI), 'Government and public sector' (GOVE), 'Health' (HEAL), 'International issues' (INTR), 'Justice, legal system and public safety' (JUST), 'Regions and cities' (REGI), 'Population and society' (SOCI), 'Science and technology'

[7] https://doi.org/10.5281/zenodo.18317554
[8] https://data.europa.eu/sparql

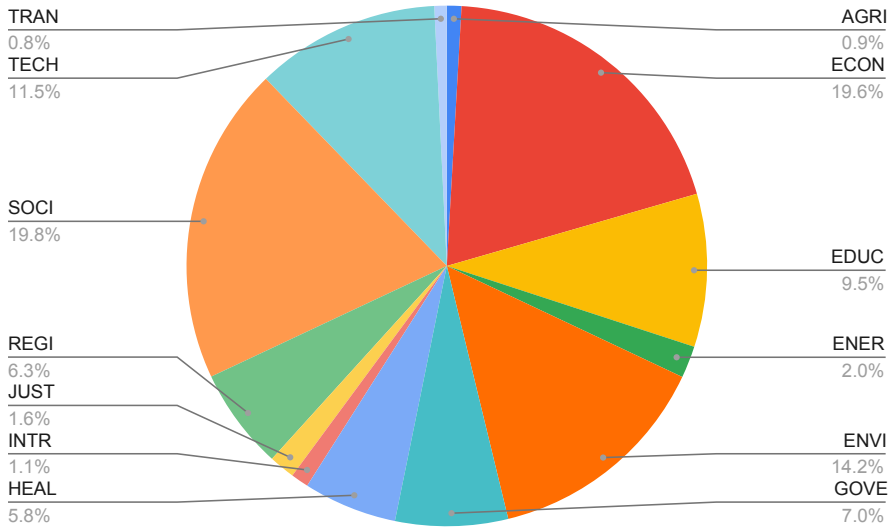


Fig. 7. Distribution of datasets across the 13 themes of the European Data Portal.

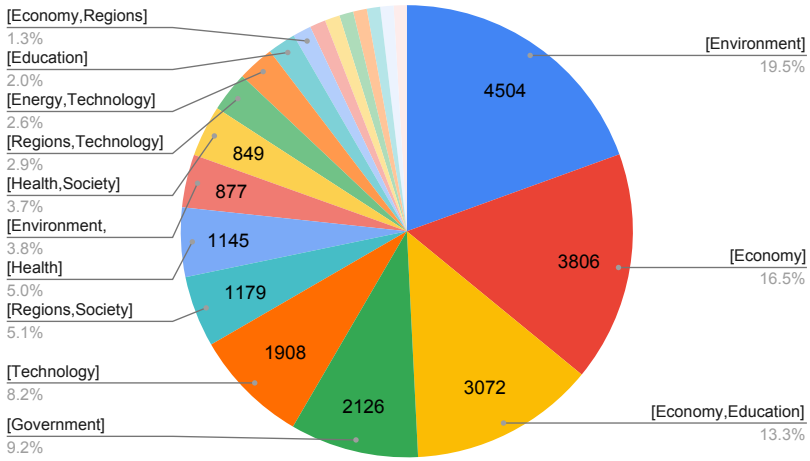


Fig. 8. Distribution of datasets with respect to the 20 most frequent combinations of themes.

(TECH), and 'Transport' (TRAN). In addition to this, as the datasets may be associated with more than one theme, the pie chart shown in Figure 8 illustrates the distribution of the datasets according to the 20 most frequent combination of themes.

Furthermore, after a manual inspection of the records we realized that a significant number of metadata records had metadata properties with text content in a different lan-

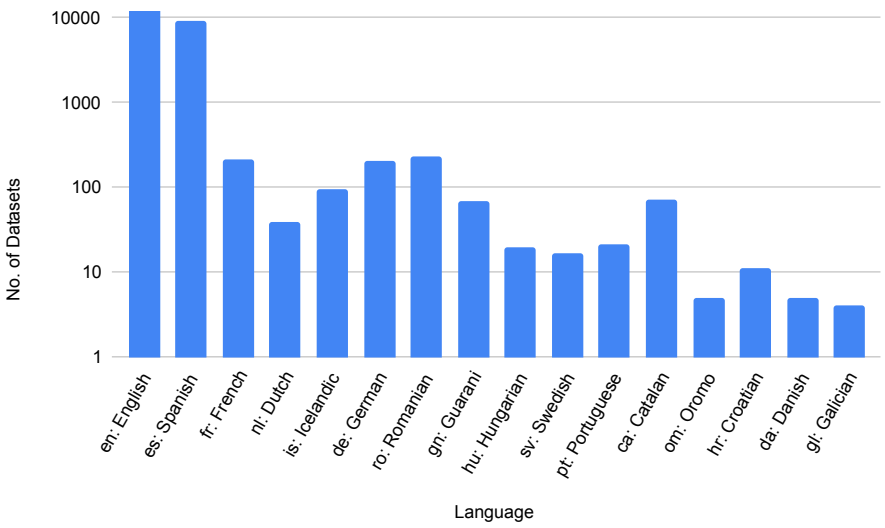


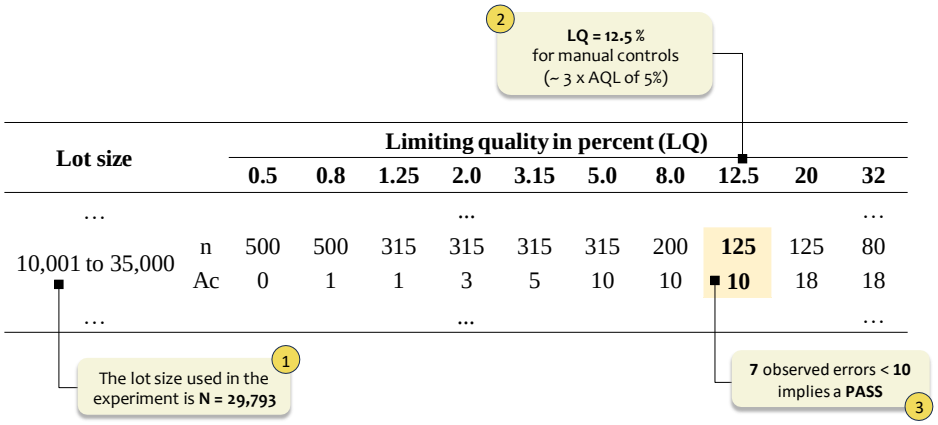
Fig. 9. The language distribution of the datasets.

guage from English. Although metadata records were specifically retrieved declaring the use of English as the language attribute for string literal values, this was not the case for many of the records. Using an API to detect the most likely source language, Figure 9 shows the distribution of languages employed in the corpus. This circumstance motivated the translation of the input text into English as a normalization process during term extraction for some experiments in Section 4.3. In a similar way, it was noticed that 18,633 words found in the input text of the corpus were not recognized as common English words, and this motivated a tailored normalization level for some experiments in Section 4.3 to remove these noise words.

4.2. Results of thematic classification correctness evaluation

Considering an AQL of 5% of errors in the thematic classification of corpus datasets, the Limiting Quality (LQ) that must be applied to a corpus that is manually inspected is thrice the AQL. As the table of ISO 2859-2 standard [24] defining the relationship between the lot size of the corpus and the selected LQ does not provide a value for 15%, the most approximate value of 12.5% must be selected.

Figure 10 describes the process followed to identify the size (n) of the sample that must be evaluated and the maximum number of errors (Ac) that can be accepted in Table A of ISO 2859-2 taking into account that our corpus consists of 29,793 datasets and we want an LQ of 12.5%. Following this, a sample of 125 records was randomly selected. The sample was then evaluated by two experts, who manually visited the dataset resources and assigned to them between one and three related themes according to the perceived content of the dataset and the text contained in title, description, and keyword properties. As indicated in section 3.1, the cases where none of the initially assigned dataset themes



Lot size		Limiting quality in percent (LQ)									
		0.5	0.8	1.25	2.0	3.15	5.0	8.0	12.5	20	32
…					…						…
10,001 to 35,000	n	500	500	315	315	315	315	200	**125**	125	80
	Ac	0	1	1	3	5	10	10	**10**	18	18
…					…						…

Fig. 10. Results of thematic classification correctness for a lot size of 29,793 records and LQ of 12.5%.

matched with one of the themes assigned by the experts were considered as errors. Upon this criterion, only 7 cases of incorrect classification were counted. As the number of errors was below the *Ac* threshold of 10 items, the quality control was passed and the thematic classification of the corpus was considered correct.

4.3. Results of automated supervised classification

This section presents the experiments performed to build models for the automated thematic annotation of datasets using the proposed approach in section 3.2 for supervised classification. Tables 1, 2 and 3 show the description of the 54 experiments that were performed considering different variants for input datasets, term extraction, feature representation and use of machine learning techniques:

- The *Input* column indicates the alternatives used for the input records. The default alternative is the use of the annotated corpus of 29,793 records (denoted as *core*). A second alternative, as proposed in section 3.2, was the incorporation of artificial datasets generated from associated themes in GEMET and UNESCO thesauri. Following this approach, we generated 686 additional records and an extended corpus of 30,479 (denoted as *extended*).
- The *Term extraction* column indicates the alternatives for term extraction: the *basic*, *translation* and *tailored* normalization levels explained in section 3.2.
- The *Feature representation* column indicates the alternatives for feature representation: the use of unigrams (*uni*); the combined use of unigrams and bigrams (*uni+bi*); and the combined use of unigrams, bigrams and trigrams (*uni+bi+tri*). The number of dimensions in the vector representation of each alternative is shown in the tables within parentheses.
- The *Classification technique* column indicates the alternatives for machine learning classification techniques (*LR*, *MNB*, or *SVM*). As indicated in section 3.2, we used the

OvR classifier to solve our multi-class classification problem. When making a prediction, all available binary classifiers are applied to the input data until one produces a confidence score high enough to be considered trustworthy. This method simplifies complex multi-class problems into binary decisions, and it enhances classification performance by focusing on differences between classes [55].

Table 1. Experiments and results for automated supervised classification: *core* input.

#	Input	Term extraction	Feature representation	Classification technique	Accuracy
1	core	basic	uni	LR	0.8881
2			(35,891)	MNB	0.7710
3				SVM	0.9365
4	core	basic	uni+bi	LR	0.84114
5			(290,600)	MNB	0.57377
6				SVM	0.8995
7	core	basic	uni+bi+tri	LR	0.8073
8			(659,880)	MNB	0.5137
9				SVM	0.8503
10	core	basic	uni	LR	0.8854
11		+ translation	(25,622)	MNB	0.7817
12				SVM	0.9355
13	core	basic	uni+bi	LR	0.8417
14		+ translation	(265,399)	MNB	0.5472
15				SVM	0.8920
16	core	basic	uni+bi+tri	LR	0.8082
17		+ translation	(619,227)	MNB	0.4969
18				SVM	0.8394

Tables 1, 2 and 3 also include a column with the accuracy obtained for each experiment. This accuracy is computed according to equation 2 taking into account the number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

SVM is the machine learning technique that performed best for all the variants incorporated in the experiments related to the input, term extraction, and feature representation. We also computed the confusion matrices for each theme. For instance, Figure 11 shows the confusion matrices for each individual theme in the best experiment, i.e., experiment 3 in Table 1.

Figure 11 also includes the precision, recall and F1 evaluation metrics according to formulas in equation (3):

$$Precision = \frac{TP}{TP + FP}; Recall = \frac{TP}{TP + FN}; F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3}$$

In addition, Figure 12 shows the curve known as the receiver operating characteristic (ROC) for experiment 3. The ROC curve is a plot of the True Positive Rate (TPR) against the False Positive Rate (FPR). Whereas TPR reflects the percentage of cases that were properly labelled as positive, FPR reflects the proportion of instances that were incorrectly

Table 2. Experiments and results for automated supervised classification: *extended* input; *basic* and *translation* normalization.

#	Input	Term extraction	Feature representation	Classification technique	Accuracy
19	extended	basic	uni	LR	0.8751
20			(30,443)	MNB	0.7654
21				SVM	0.9226
22	extended	basic	uni+bi	LR	0.8288
23			(280,834)	MNB	0.5656
24				SVM	0.8827
25	extended	basic	uni+bi+tri	LR	0.7959
26			(645,748)	MNB	0.4992
27				SVM	0.8325
28	extended	basic	uni	LR	0.8721
29		+ translation	(26,497)	MNB	0.7623
30				SVM	0.9217
31	extended	basic	uni+bi	LR	0.8298
32		+ translation	(273,849)	MNB	0.5346
33				SVM	0.8754
34	extended	basic	uni+bi+tri	LR	0.7935
35		+ translation	(638,605)	MNB	0.4769
36				SVM	0.8254

Table 3. Experiments and results for automated supervised classification: *extended* input; *tailored* normalization.

#	Input	Term extraction	Feature representation	Classification technique	Accuracy
37	extended	basic	uni	LR	0.8715
38		+ tailored	(17,371)	MNB	0.7737
39				SVM	0.9152
40	extended	basic	uni+bi	LR	0.8248
41		+ tailored	(222,559)	MNB	0.5242
42				SVM	0.8720
43	extended	basic	uni+bi+tri	LR	0.7909
44		+ tailored	(529,092)	MNB	0.4647
45				SVM	0.8201
46	extended	basic	uni	LR	0.8677
47		+ translation	(12,906)	MNB	0.7696
48		+ tailored		SVM	0.9142
49	extended	basic	uni+bi	LR	0.8236
50		+ translation	(215,844)	MNB	0.4971
51		+ tailored		SVM	0.8632
52	extended	basic	uni+bi+tri	LR	0.7886
53		+ translation	(524,646)	MNB	0.4433
54		+ tailored		SVM	0.8090

classified as positive (see formulas in equation 4). It can be observed that the area under the curve (AUC) of the ROC curve is close to the maximum value for practically all of the themes, which demonstrates that the configuration of the SVM experiment has a high probability to assign correctly the theme of a dataset.

$$TPR = \frac{TP}{TP + FN}; FPR = \frac{FP}{FP + TN} \tag{4}$$

4.4. Results of theme prediction

This section presents the results of the approach proposed in section 3.3 to predict automatically the closest theme according to the similarity between the word/sentence em-

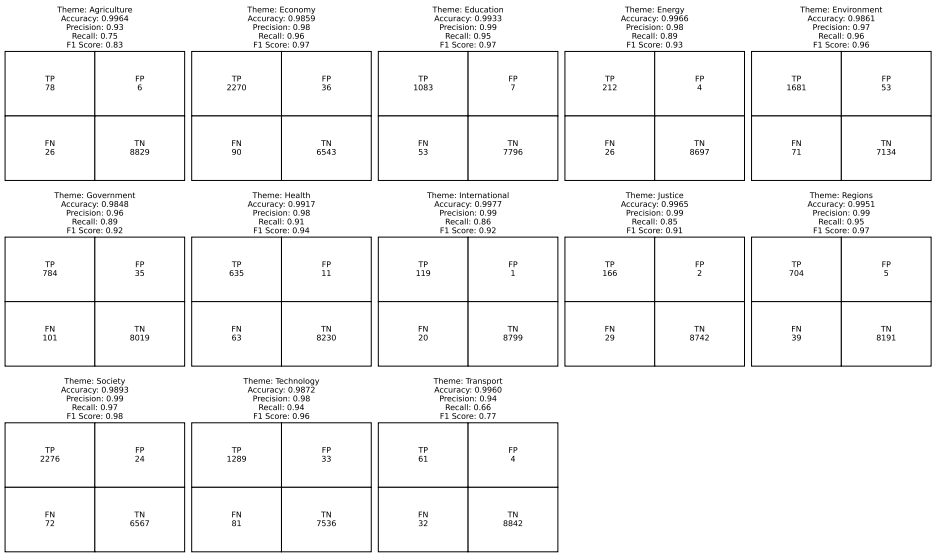


Fig. 11. Confusion matrices for all the themes in experiment 3 of Table 1 (*core* input, *basic* normalization, unigram features, SVM, overall accuracy of 93.65%)

beddings of the metadata content and the definition of the European Data themes. It is an unsupervised approach to predict themes. Table 4 shows the description of the 7 experiments that were performed to automatically assign themes to the datasets in our corpus considering different variants for property selection, term extraction, and feature representation:

- The *Dataset property selection* column indicates the alternatives for the selection of metadata properties describing the datasets as proposed in s̃ection 3.3: the concatenation of three properties (*title+description+keywords*) or just the concatenation of title and description (*title+description*). It must be noted that the property selection for themes is not detailed because it is maintained in all the experiments: the input text is the concatenation of the preferred label and definition of each theme in English.
- The *Term extraction* column indicates the alternatives for term extraction explained in section 3.3: *basic* normalization and the additional process of *stop word removal* in some cases.
- The *Feature representation* column shows the alternatives that have been used for feature representation as proposed in section 3.3: *GloVe sum* indicates the use of GloVe word embeddings and the representation of the full text as the sum of the embeddings of each word in the text; *GloVe average* indicates the use of GloVe word embeddings the representation of the full text as the sum of the embeddings of each word in the text; *BERT* indicates the use of BERT sentence embeddings; and *HuggingFace* indicates the use of HuggingFace transformers for sentence embeddings.

In order to have an orientation about the appropriateness of the predicted themes by the different experiments, we compared the top three themes (ranked by decreasing cosine

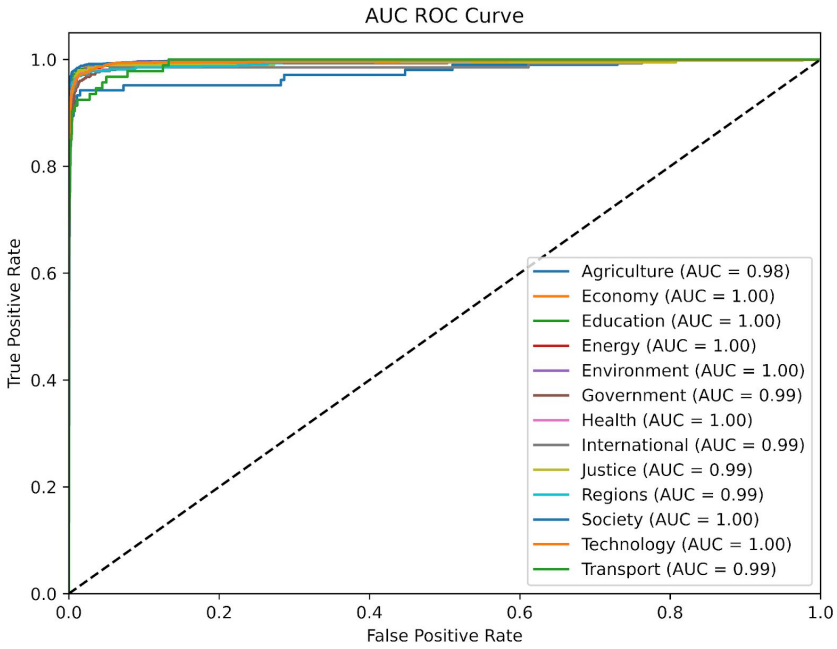


Fig. 12. ROC curve for all the themes in experiment 3 of Table 1 (*core* input, *basic* normalization, unigram features, SVM, overall accuracy of 93.65%)

Table 4. Experiments and results for theme prediction

#	Dataset property selection	Term Extraction	Feature Representation	Agreement score		
				Top 1	Top 2	Top 3
1	title + description + keywords	basic	GloVe sum	0.4127	0.5394	0.6536
2	title + description + keywords	basic	GloVe average	0.4397	0.5770	0.6865
3	title + description + keywords	basic + stop-word removal	GloVe average	0.4680	0.6186	0.7253
4	title + description	basic	BERT	0.2480	0.3360	0.3920
5	title + description + keywords	basic	BERT	0.1908	0.3132	0.4109
6	title + description	basic	HuggingFace	0.3920	0.6320	0.7120
7	title + description + keywords	basic	HuggingFace	0.5023	0.6734	0.7456

similarity distance) with the original dataset themes assigned in the annotated corpus. Table 4 includes an agreement score for the top 1, top 2 and top 3 themes. This agreement score measures the proportion of matches between the top 1/2/3 themes and the assigned themes in the corpus. For instance, if a dataset was originally annotated with the "society" theme, and "society" is the third more relevant theme assigned, this is a match for the top

3 agreement score. Equation 5 shows the formula for computing the agreement score of a $corpus$ and top n predicted themes: $themes(d)$ stands for the function that returns the themes assigned to a dataset d in the $corpus$; $predicted\_themes(d, n)$ stands for the function that returns the top n predicted themes of a dataset d; and $|corpus|$ is the number of datasets in the corpus.

$$Agreement\_score(corpus, n) = \frac{\sum_{d \in corpus} \begin{cases} 1, & \text{if } themes(d) \cap predicted\_themes(d, n) \neq \emptyset \\ 0, & \text{otherwise} \end{cases}}{|corpus|} \tag{5}$$

It can be observed that the best agreement score is obtained with the HuggingFace Transformer for sentence embeddings in experiment 7 and comparing the 3 best ranked themes with the original dataset themes.

5. Discussion

This section discusses the experimental results explicitly in relation to the research questions (RQs) formulated in the Introduction section and highlights how the proposed framework for thematic annotation of OGD addresses each of them. The feasibility of our thematic annotation framework was tested through a series of experiments using a corpus of metadata records which is a representative sample of the current metadata describing OGD in Europe. In the first part of the experiments, we evaluated the correctness of the existing thematic annotations. Although we assessed that the thematic classification correctness of the annotated corpus had an acceptable quality with less than 5% of errors, it must be acknowledged that in many cases the original theme of the inspected datasets for quality control was not the first option in the proposed list of up to three themes assigned by the experts doing the evaluation. This observation is consistent with prior studies highlighting how inconsistencies in metadata structuring and categorization negatively affect the usability and usefulness of open government data [5]. This claim motivates the need for a framework that assists in the thematic annotation relying on the training with a big corpus of annotated datasets, where the biases are minimized. Overall, the results obtained for RQ1 indicate that existing thematic annotations in large OGD portals are generally acceptable but not free from inconsistencies, thereby justifying the need for systematic and scalable support mechanisms such as the proposed framework.

The second research question (RQ2) investigates the extent to which new datasets can be automatically classified when a properly annotated corpus is available. As a second part of the experiments, we analyzed the feasibility of different machine learning techniques to build models for automatic thematic annotation of datasets. In the experiments we cleaned the DCAT-based metadata text by applying several text processing techniques, which also included the translation of false English text and the removal of non-common English terms. With respect to feature representation, we also tested the combined use of unigrams, bigrams and trigrams. Supervised classification techniques such as Logistic Regression and Naive Bayes, as well as Support Vector Machines (SVM), showed effectiveness in classifying datasets with themes using titles, descriptions and keywords, being SVM the technique having the highest accuracy of 93.65%. These results provide a

clear and positive answer to RQ2, demonstrating that supervised machine learning techniques particularly SVM can be effectively integrated into the proposed framework to support large-scale automatic thematic annotation when high-quality annotated metadata are available. Moreover, the results are consistent with similar works in related domains where SVM has also provided a high accuracy for supervised classification [21].

The third research question (RQ3) addresses the issues about the absence of a properly annotated corpus and how can relevant themes be assigned to a dataset based solely on free-text metadata. The final part of our experiments also considered the possibility of not having an annotated corpus or not counting on a perfect annotated corpus with themes. In this case we proposed a representation of texts derived from metadata and theme descriptions in terms of word or sentence embeddings. To predict the themes closer to a dataset, we computed the cosine distance between the embedding representations of the dataset and the candidate themes. After doing experiments with the same sample of datasets extracted from *data.europa.eu* and different techniques for word embeddings (GloVe) and sentence embeddings (BERT and HuggingFace Transformers), we concluded that HuggingFace Transformers were the best approach. The predicted themes have a high agreement score (74.56%) with respect to the original themes assigned in the European data portal. Although the obtained agreement score of 0.7456 is lower than the accuracy obtained with the best experiment for classification models (0.9365), these numbers are not comparable. In the component for thematic prediction of our framework for thematic annotation, we are assuming that the initial thematic annotation is not perfect (or not existent) and we try to identify the closest theme according to the similarities of the language models used to generate the embeddings of dataset metadata and theme descriptions. In some cases, the definition proposed by the European Union [48] for a theme consists of a reduced number of words (sentences) and may not encompass all the possible aspects that the datasets associated with this theme may cover. For instance, the 'Health' theme is defined in just three sentences.[9] Larger texts would generate an embedding vector representation with a better alignment with all the aspects covered by a dataset theme. The experts that evaluated the thematic classification correctness assessed that there were not more than 5% of errors in the classification, but their manual annotation of themes was not constrained by the short definitions of themes. The results obtained for RQ3 show that embedding-based approaches constitute a viable alternative within the framework when annotated corpora are missing or unreliable, thereby increasing the applicability of the proposed thematic annotation framework. The obtained results are coherent with the reported experiments of similar works such as the one proposed by Huseynov et al. [22] for a dataset recommender, which also employed embedding-based representations of metadata and cosine similarity to identify the closer datasets.

6. Conclusions

This paper has presented a framework for the thematic annotation of OGD, which has been tested against a representative sample of 29,793 datasets from *data.europa.eu*, a portal that aggregates datasets (and their associated metadata) harvested from both the

[9] *"This concept identifies datasets covering the domain of health. Health is a state of physical, mental and social well-being in which disease and infirmity are absent. Dataset examples: COVID-19 Coronavirus data; European Cancer Information System."*

member states of the European Union and the European institutions. With respect to the research questions formulated in this study, the results allow us to draw the following conclusions. First, in response to RQ1, we showed that while existing thematic annotations in large OGD portals show an overall acceptable level of correctness, they also present inconsistencies and subjectivity, thus motivating the need for systematic support mechanisms. Second, addressing RQ2, we demonstrated that supervised machine learning techniques, especially SVM, can accurately classify new datasets when a well annotated metadata corpus is available. Finally, in response to RQ3, we confirmed that embedding-based approaches using free-text metadata and theme descriptions provide an applicable solution when annotated corpora are missing. Together, these results validate the design of the proposed framework as a comprehensive and flexible solution for thematic annotation in heterogeneous OGD environments.

6.1. Theoretical Contributions

From a theoretical perspective, this work advances the understanding of thematic annotation in the context of open government data by conceptualizing it as a structured and multi-component process. The proposed framework integrates evaluation of existing annotations, supervised classification, and embedding-based thematic prediction into a unified model, offering a systematic view of how different annotation strategies can be combined depending on data availability and quality. By explicitly addressing multiple annotation scenarios, this study contributes to the literature on metadata quality, semantic enrichment, and data findability in OGD ecosystems.

6.2. Practical Contributions

From a practical standpoint, the proposed framework provides actionable guidance for practitioners and open data portal operators seeking to improve dataset discoverability and consistency of thematic categorization. The framework can be integrated into the dataset ingestion pipelines of widely used open data platforms such as CKAN, DKAN, or Socrata using metadata models based on DCAT where general properties like title, description, keywords, and themes are available. The supervised classification algorithms can be trained to facilitate the automatic annotation of new inserted metadata records. The only requirement for customizing the framework to metadata in other languages is to adjust the term extraction libraries for a satisfactory performance of tokenization, stop removal or other text pre-processing steps in specific languages. In the case of unsupervised classification for theme prediction, we would just need to select pre-trained models for word/sentence embeddings (e.g., Glove, BERT, . . .) in specific languages. By reducing reliance on manual annotation and mitigating subjectivity, the framework has the potential to improve the usability of open data portals and facilitate more efficient dataset discovery for diverse user groups.

6.3. Limitations

This study is also subject to several limitations. Although *data.europa.eu* stands as one of the largest open data government portals and serves as a hub for the national OGD portals

of the European countries, it is crucial to recognize that the categorization of datasets may heavily reflect the biases of the entities responsible for publishing them. The performance of the framework may be influenced by the selection of the vocabulary for data themes because the appropriateness of their titles (preferred labels) and definitions is essential for the assessment of the thematic classification correctness of the annotated corpus (later used in supervised classification techniques) and the approach proposed for theme prediction, which relies on the generation of an embedding-based representation of each theme definition. Exploring the relationship and compatibility of thematic classification schemes employed in OGD portals across other regions [8] could enhance the representativeness and generalization of automated thematic classification algorithms.

In addition, the quality of metadata presents another significant constraint. The prevalence of datasets nominally labelled in English but containing text in other languages exemplifies the noise inherent in the training data. Consequently, sensitivity to such noise emerges as a pertinent consideration in the algorithmic approach to the thematic classification of datasets.

Last, it must be observed that our framework has not been integrated and tested within the scope of an open data portal with end users. Our framework is not aimed at being directly executed by end users interacting with open data portals, but to be integrated during the ingestion process of datasets in a data portal. In order to simulate the thematic annotation during this ingestion process, this work reports experiments whose results have been evaluated in terms of relevance measures, which are employed in the information retrieval discipline to estimate user satisfaction. However, we acknowledge that techniques like A/B testing [50] could be used to verify with end users if an open data portal incorporating this innovation during the ingestion process is better accepted than the portal without the innovation. For instance, we could compare the number of clicks on the first hits returned by both portals with thematic searches.

6.4. Future Research Directions

Building on the findings and limitations of this study, several avenues for future research emerge. First, we would like to explore if the information related to the application schema of the different distributions of datasets can help us to improve the automatic thematic classification of datasets. Available distributions in machine readable formats such as CSV or RDF can provide in some cases meaningful names of thematic attributes of the dataset content. Even in the case of RDF (graph data), these attributes are usually selected from well-known vocabularies, and this may be used to infer links with the themes that can be assigned automatically. Second, we could also explore alternative approaches to unsupervised classification for theme prediction based on the use of keyword extraction techniques [2] and see whether the extracted keywords align with the salient keywords of theme definitions. Last, the impact of data policies on thematic annotation practices and the user experience in accessing and utilizing annotated data could be more deeply investigated to understand how regulations influence the effectiveness of open data ecosystems.

References

1. Ahmed, U.: Reimagining open data ecosystems: a practical approach using AI, CI, and knowledge graphs. In: BIR Workshops. pp. 235–249 (2023)

2. Ahmed, U., Alexopoulos, C., Piangerelli, M., Polini, A.: BRYT: Automated keyword extraction for open datasets. Intelligent Systems with Applications 23, 200421 (2024)
3. Alexopoulos, C., Loukis, E., Charalabidis, Y.: A methodology for determining the value generation mechanism and the improvement priorities of open government data systems. Computer Science and Information Systems 13(1), 237–258 (2016)
4. Alexopoulos, C., Spiliotopoulou, L., Charalabidis, Y.: Open data movement in Greece: a case study on open government data sources. In: Proceedings of the 17th Panhellenic Conference on Informatics. pp. 279–286 (2013)
5. Ansari, B., Barati, M., Martin, E.G.: Enhancing the usability and usefulness of open government data: A comprehensive review of the state of open government data visualization research. Government Information Quarterly 39(1), 101657 (2022)
6. Arlotta, L., Crescenzi, V., Mecca, G., Merialdo, P.: Automatic annotation of data extracted from large web sites. In: International Workshop on the Web and Databases. pp. 7–12 (2003)
7. Attard, J., Orlandi, F., Scerri, S., Auer, S.: A systematic review of open government data initiatives. Government information quarterly 32(4), 399–418 (2015)
8. Bogdanović, M., Gligorijević, M.F., Veljković, N., Puflović, D., Stoimenov, L.: Cross-portal metadata alignment–connecting open data portals through means of formal concept analysis. Information Sciences 637, 118958 (2023)
9. Carducci, G., Leontino, M., Radicioni, D.P., Bonino, G., Pasini, E., Tripodi, P.: Semantically aware text categorisation for metadata annotation. In: Italian Research Conference on Digital Libraries. pp. 315–330. Springer (2019)
10. Davies, T., Walker, S.B., Rubinstein, M., Perini, F.: The state of open data: Histories and horizons. African Minds (2019)
11. Dekkers, M., Kotoglou, S., Nelson, C., Pellegrino, M., Hohn, N., Peristeras, V.: StatDCAT-AP, a common layer for the exchange of statistical metadata in open data portals. In: 6th International Workshop on Semantic Statistics co-located with the 17th International Semantic Web Conference (2016)
12. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of North American Association for Computational Linguistics. vol. 1, p. 2 (2019)
13. Díaz-Corona, D., Lacasta, J., Latre, M.Á., Zarazaga-Soria, F.J., Nogueras-Iso, J.: Profiling of knowledge organisation systems for the annotation of linked data cultural resources. Information Systems 84, 17–28 (2019)
14. Ellen, J.S., Graff, C.A., Ohman, M.D.: Improving plankton image classification using context metadata. Limnology and Oceanography: Methods 17(8), 439–461 (2019)
15. Enríquez-Reyes, R., Cadena-Vela, S., Fuster-Guilló, A., Mazón, J.N., Ibáñez, L.D., Simperl, E.: Systematic mapping of open data studies: Classification and trends from a technological perspective. IEEE Access 9, 12968–12988 (2021)
16. European Commission: DCAT Application profile for data portals in Europe, DCAT-AP Version 2.1.0 (2021), https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/dcat-application-profile-data-portals-europe/release/210
17. Evans, A.M., Campos, A.: Open government initiatives: Challenges of citizen participation. Journal of policy analysis and management pp. 172–185 (2013)
18. Freire, J., Fan, G., Feuer, B., Koutras, C., Liu, Y., Peña, E., Santos, A.S., Silva, C.T., Wu, E.: Large language models for data discovery and integration: Challenges and opportunities. IEEE Data Eng. Bull. 49(1), 3–31 (2025)
19. Haunss, S., Kuhn, J., Padó, S., Blessing, A., Blokker, N., Dayanik, E., Lapesa, G.: Integrating manual and automatic annotation for the creation of discourse network data sets. Politics and Governance 8(2), 326–339 (2020)

20. Hudon, A., Beaudoin, M., Phraxayavong, K., Dellazizzo, L., Potvin, S., Dumais, A.: Use of automated thematic annotations for small data sets in a psychotherapeutic context: systematic review of machine learning algorithms. JMIR mental health 8(10), e22651 (2021)
21. Hudon, A., Beaudoin, M., Phraxayavong, K., Dellazizzo, L., Potvin, S., Dumais, A.: Implementation of a machine learning algorithm for automated thematic annotations in avatar: A linear support vector classifier approach. Health Informatics Journal 28(4), 14604582221142442 (2022)
22. Huseynov, R., Nikiforova, A., Symeonidis, D., Duenas-Cid, D.: May the Data Be with You: Towards an AI-Powered Semantic Recommender for Unlocking Dark Data. In: International Conference on Electronic Government. Springer (2025)
23. Ibáñez, L.D., Millard, I., Glaser, H., Simperl, E.: An assessment of adoption and quality of linked data in European open government data. In: International Semantic Web Conference. pp. 436–453. Springer (2019)
24. International Organization for Standardization (ISO): Sampling Procedures for Inspection by Attributes–Part 2: Sampling Plans Indexed by Limiting Quality (LQ) for Isolated Lot Inspection, Standard ISO 2859-2:1985 (1985)
25. Janssen, M., Charalabidis, Y., Zuiderwijk, A.: Benefits, adoption barriers and myths of open data and open government. Information systems management 29(4), 258–268 (2012)
26. de Juana-Espinosa, S., Luján-Mora, S.: Open government data portals in the European union: A dataset from 2015 to 2017. Data in brief 29, 105156 (2020)
27. Kaldeli, E., Menis-Mastromichalakis, O., Bekiaris, S., Ralli, M., Tzouvaras, V., Stamou, G.: CrowdHeritage: crowdsourcing for improving the quality of cultural heritage metadata. Information 12(02), 64 (2021)
28. Kirstein, F., Dittwald, B., Dutkowski, S., Glikman, Y., Schimmler, S., Hauswirth, M.: Linked data in the european data portal: A comprehensive platform for applying DCAT-AP. In: International Conference on Electronic Government. pp. 192–204. Springer (2019)
29. Kliimask, K., Nikiforova, A.: TAGIFY: LLM-powered tagging interface for improved data findability on OGD portals. In: 2024 Fifth International Conference on Intelligent Data Science Technologies and Applications (IDSTA). pp. 18–27. IEEE (2024)
30. Kubler, S., Robert, J., Neumaier, S., Umbrich, J., Le Traon, Y.: Comparison of metadata quality in open data portals using the analytic hierarchy process. Government Information Quarterly 35(1), 13–29 (2018)
31. Li, B., Zhou, H., He, J., Wang, M., Yang, Y., Li, L.: On the sentence embeddings from pre-trained language models. In: Webber, B., Cohn, T., He, Y., Liu, Y. (eds.) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 9119–9130. Association for Computational Linguistics, Online (Nov 2020)
32. Lněnička, M., Machova, R., Volejníková, J., Linhartová, V., Knezackova, R., Hub, M.: Enhancing transparency through open government data: The case of data portals and their features and capabilities. Online Information Review 45(6), 1021–1038 (2021)
33. Lnenicka, M., Nikiforova, A., Luterek, M., Milic, P., Rudmark, D., Neumaier, S., Kević, K., Zuiderwijk, A., Bolívar, M.P.R.: Understanding the development of public data ecosystems: From a conceptual model to a six-generation model of the evolution of public data ecosystems. Telematics and informatics p. 102190 (2024)
34. Luna-Reyes, L.F., Bertot, J.C., Mellouli, S.: Open government, open data and digital government. Government Information Quarterly 31(1), 4–5 (2014)
35. Martín-Chozas, P., Montiel-Ponsoda, E., Rodríguez-Doncel, V.: Language resources as linked data for the legal domain. In: Knowledge of the Law in the Big Data Age, pp. 170–180. IOS Press (2019)
36. Miles, A., Brickley, D.: SKOS Core Guide. W3C Working Draft 2 November 2005 (2005), https://www.w3.org/TR/swbp-skos-core-guide

37. Mohamed, M., Pillutla, S., Tomasi, S.: Extraction of knowledge from open government data: The knowledge iterative value network framework. VINE Journal of Information and Knowledge Management Systems 50(3), 495–511 (2020)
38. Morville, P., Rosenfeld, L.: Information Architecture for the World Wide Web. O'Reilly Media, Inc., Canada (2015)
39. Mylonas, P., Voutos, Y., Sofou, A.: A collaborative pilot platform for data annotation and enrichment in viticulture. Information 10(4), 149 (2019)
40. Ni, J., Hernandez Abrego, G., Constant, N., Ma, J., Hall, K., Cer, D., Yang, Y.: Sentence-T5: Scalable Sentence Encoders from Pre-trained Text-to-Text Models. In: Findings of the Association for Computational Linguistics: ACL 2022. pp. 1864–1874. Association for Computational Linguistics, Dublin, Ireland (May 2022)
41. Nikiforova, A., McBride, K.: Open government data portal usability: A user-centred usability analysis of 41 open government data portals. Telematics and Informatics 58, 101539 (2021)
42. Nogueras-Iso, J., Lacasta, J., Ureña-Cámara, M.A., Ariza-López, F.J.: Quality of metadata in open data portals. IEEE Access 9, 60364–60382 (2021)
43. Nogueras-Iso, J., Latre, M.Á., Bejar, R., Muro-Medrano, P.R., Zarazaga-Soria, F.J.: A model driven approach for the development of metadata editors, applicability to the annotation of geographic information resources. Data & Knowledge Engineering 81, 118–139 (2012)
44. Paterne Chokki, A., Alexopoulos, C., Matheus, R., Saxena, S., Frénay, B., Vanderose, B.: Do open government data (OGD) portals show signs of knowledge management (KM) practices?: an empirical investigation. Technology Analysis & Strategic Management 36(12), 4829–4844 (2024)
45. Pavia, S., Piraino, N., Islam, K., Pyayt, A., Gubanov, M.N.: Hybrid metadata classification in large-scale structured datasets. Journal of Data Intelligence 3(4), 460–473 (2022)
46. Peng, Y., Wu, Z., Jiang, J.: A novel feature selection approach for biomedical data classification. Journal of Biomedical Informatics 43(1), 15–23 (2010)
47. Petrillo, M., Baycroft, J.: Introduction to manual annotation. Fairview research pp. 1–7 (2010)
48. Publications Office of the European Union: Data theme authority table (2022), https://op.europa.eu/s/zBx4
49. Publications Office of the European Union: Metadata quality assessment methodology. Online (2025), https://data.europa.eu/mqa/methodology?locale=en, accessed: 2025-06-06
50. Quin, F., Weyns, D., Galster, M., Silva, C.C.: A/B testing: A systematic literature review. Journal of Systems and Software 211, 112011 (2024)
51. Salih, A.Q.M.: Towards from manual to automatic semantic annotation: based on ontology elements and relationships. International Journal of Web & Semantic Technology 4(2), 21 (2013)
52. Sasse, J., Darms, J., Fluck, J.: Semantic metadata annotation services in the biomedical domain—a literature review. Applied Sciences 12(2), 796 (2022)
53. Shah, S.I.H., Peristeras, V., Magnisalis, I.: A conceptual framework for the government big data ecosystem ('datagov. eco'). Data & Knowledge Engineering p. 102348 (2024)
54. Simonofski, A., Nikiforova, A., Lnenicka, M., Bono Rossello, N.: Artificial intelligence as a catalyzer for open government data ecosystems: A typological theory approach (2025)
55. Tao, W., Yongjia, J., Xiangsheng, R.: A novel two-level one-vs-rest classifier. In: 2019 2nd International Conference on Information Systems and Computer Aided Education (ICISCAE). pp. 645–648. IEEE (2019)
56. Tuarob, S., Pouchard, L.C., Noy, N.F., Horsburgh, J.S., Palanisamy, G.: ONEMercury: Towards automatic annotation of environmental science metadata. In: Proceedings of the Second International Workshop on Linked Science 2012 - Tackling Big Data, Boston, MA, USA,. CEUR Workshop Proceedings, vol. 951 (2012)

57. Ureña-Cámara, M.A., Nogueras-Iso, J., Lacasta, J., Ariza-López, F.J.: A method for checking the quality of geographic metadata based on ISO 19157. International Journal of Geographical Information Science 33(1), 1–27 (2019)
58. Vandenbussche, P.Y., Vatant, B.: Metadata recommendations for linked open data vocabularies. Version 1, 2011–12 (2011)
59. Verberne, S., D'hondt, E., Van den Bosch, A., Marx, M.: Automatic thematic classification of election manifestos. Information Processing & Management 50(4), 554–567 (2014)
60. W3C: Data Catalog Vocabulary (DCAT). W3C Recommendation 16 January 2014 (2014), https://www.w3.org/TR/vocab-dcat/
61. Wentzel, B., Kirstein, F., Jastrow, T., Sturm, R., Peters, M., Schimmler, S.: An extensive methodology and framework for quality assessment of DCAT-AP datasets. In: International Conference on Electronic Government. pp. 262–278. Springer (2023)
62. Wilkinson, M., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., Bonino da Silva Santos, L.O., Bourne, P., Bouwman, J., Brookes, A., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C., Finkers, R., Mons, B.: The FAIR guiding principles for scientific data management and stewardship. Scientific data 3(1), 1–9 (2016)
63. Wu, M., Brandhorst, H., Marinescu, M.C., Lopez, J.M., Hlava, M., Busch, J.: Automated metadata annotation: What is and is not possible with machine learning. Data Intelligence 5(1), 122–138 (2023)
64. Yimam, S.M., Biemann, C., de Castilho, R.E., Gurevych, I.: Automatic annotation suggestions and custom annotation layers in webanno. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 91–96 (2014)
65. Zhang, H., Liu, Y., Santos, A., Freire, J., et al.: Autoddg: Automated dataset description generation using large language models. arXiv preprint arXiv:2502.01050 (2025)

Abdul Aziz received the bachelor's degree in computer science from the COMSATS Institute of Information Technology, Lahore, Pakistan, in 2013, and the master's degree in computer science from the National University of Computer and Emerging Sciences, Karachi, Pakistan, in 2018. In 2025 he defended his Ph.D. degree in Computer Science at the University of Zaragoza (Advanced Information Systems Laboratory of the Aragon Institute of Engineering Research), Spain, about the use of feedback mechanisms to promote the Inclusiveness of open government data portals. During his Ph.D. he was an Early Stage Researcher in the ODECO project, a Horizon 2020 Marie Skłodowska-Curie Innovative Training Network (H2020-MSCA-ITN-2020), where he contributed to advancing research on open data ecosystems, user engagement, and data-driven innovation. He is currently working as an AI & Data Consultant at PQNO?, where he applies advanced artificial intelligence and data-driven methodologies to support innovation, strategic decision-making, and digital transformation across diverse domains.

Mohsan Ali is a researcher at the University of the Aegean. He was awarded a Marie Skłodowska-Curie Innovative Training Network (H2020-MSCA-ITN-2020) scholarship in 2021 to pursue his PhD in Greece, focusing on open data ecosystems. His current research centers on the technical interoperability of open data within the Information Systems Laboratory, as part of the ODECO-funded project. His expertise includes open data, data interoperability, data science, natural language processing, and artificial intelligence. In addition, he has specialized in deep learning, a skill developed through his academic and professional training. He holds a Master's degree in Computer Science (MSCS) with

distinction and was awarded a Gold Medal from Air University, Islamabad, Pakistan. He also earned his Bachelor's degree from Pir Mehr Ali Shah Arid Agriculture University, Rawalpindi, Pakistan.

Dagoberto Jose Herrera-Murillo received the bachelor's degree in Business Informatics from Tecnológico de Monterrey and the joint master's degree in Big Data Management from Université Libre de Bruxelles, Universitat Politècnica de Catalunya (BarcelonaTech), and Eindhoven University of Technology. In 2025 he defended his Ph.D. degree in Computer Science at the University of Zaragoza (Advanced Information Systems Laboratory of the Aragon Institute of Engineering Research), Spain, about the evaluation of user interfaces of open data portals . During his PhD he was an Early Stage Researcher for the ODECO project, a Horizon 2020 Marie Sklodowska-Curie Innovative Training Network (H2020-MSCA-ITN-2020).

Maria Ioanna Maratsi is a researcher at the University of the Aegean (Department of Information and Communication Systems Engineering) and PhD candidate in the area of semantic interoperability and open data. She is a graduate (BSc) of Computer Science and Telecommunications from University of Piraeus, Greece, and alumna of the MSc in Information Security of Stockholm University (Department of Computer and Systems Sciences - DSV), Sweden. Ioanna was also a member of the Systems Analysis and Security Unit of Stockholm University. During her PhD she was an Early Stage Researcher for the ODECO project, a Horizon 2020 Marie Sklodowska-Curie Innovative Training Network (H2020-MSCA-ITN-2020). Her academic interests include open and linked data, interoperability, knowledge graph engineering, data privacy, digital forensics, AI ethics, and multidisciplinary approaches in information technology.

Francisco Javier Lopez-Pellicer received the M.S. and Ph.D. degrees in computer engineering from the University of Zaragoza. In 2004, he started his research with the Advanced Information Systems Laboratory, University of Zaragoza (Spain). Currently, he is an Associate Professor of computer science at the University of Zaragoza. Over the past ten years, his professional career has been linked to open data initiatives and spatial data infrastructures. Within this context, he has coauthored numerous publications in books, journals or conference proceedings; and has collaborated in several R+D projects. His research interests include open data infrastructures, service-based geographic information systems, and various information systems.

Javier Nogueras-Iso received the M.S. and Ph.D. degrees in computer science from the University of Zaragoza, Spain. In 1998, he started his research with the Advanced Information Systems Laboratory, University of Zaragoza (Spain), where he is currently a Full Professor of computer science. From 2011 to 2017, he was the Director of the Catedra Logisman on Technological Document Management. From 2015 to 2019, he was the Associate Director of the Aragon Institute of Engineering Research (I3A). His research interests include information retrieval and semantic web technologies applied to different domains, although with a special emphasis on geographic information infrastructures.

Received: October 29, 2025; Accepted: April 6, 2026.