International Journal on Digital Libraries
https://doi.org/10.1007/s00799-021-00302-1

Automatically evaluating the quality of textual descriptions in cultural heritage records

Matteo Lorenzini[1,2] · **Marco Rospocher**[3] · **Sara Tonelli**[2]

Received: 10 July 2020 / Revised: 26 March 2021 / Accepted: 2 April 2021

Abstract
Metadata are fundamental for the indexing, browsing and retrieval of cultural heritage resources in repositories, digital libraries and catalogues. In order to be effectively exploited, metadata information has to meet some quality standards, typically defined in the collection usage guidelines. As manually checking the quality of metadata in a repository may not be affordable, especially in large collections, in this paper we specifically address the problem of automatically assessing the quality of metadata, focusing in particular on textual descriptions of cultural heritage items. We describe a novel approach based on machine learning that tackles this problem by framing it as a binary text classification task aimed at evaluating the accuracy of textual descriptions. We report our assessment of different classifiers using a new dataset that we developed, containing more than 100K descriptions. The dataset was extracted from different collections and domains from the Italian digital library "Cultura Italia" and was annotated with accuracy information in terms of compliance with the cataloguing guidelines. The results empirically confirm that our proposed approach can effectively support curators (F1 $\sim$0.85) in assessing the quality of the textual descriptions of the records in their collections and provide some insights into how training data, specifically their size and domain, can affect classification performance.

Keywords Metadata quality · Digital libraries · Cultural heritage · Natural language processing · Machine learning

1 Introduction

In the last years, the number of digital repositories in the cultural heritage domain has remarkably increased. These digital repositories are indexed by means of descriptive metadata (i.e. data that give information about the content of a collection), representing the backbone through which users can navigate information and improve their knowledge of specific topics, also reusing data coming from external sources [6,16]. For this reason, managing and maintaining correct information in metadata throughout their entire lifecycle plays a fundamental role [30]. However, the process of quality control still lacks a clear definition and workflow. This has several implications, including the impossibility of introducing systematic approaches to its automatic measurement and enhancement [14].

Defining what metadata quality is and how it should be measured is a very challenging task. No consensus has been reached on this concept yet. This is indeed a multidimensional and context-specific notion [37], whose definition and quantification change depending on the function of the digital archive and domain [15,37]. Building upon past works, we therefore adopt an operational definition of metadata quality, considering it as a way to measure how much the information describing a cultural heritage object supports a given purpose [31].

A number of metadata quality criteria have been suggested to guide metadata management and evaluation. One of the best-known frameworks for metadata quality has been proposed by Bruce and Hillmann [5] and includes seven qualitative dimensions to measure metadata quality: *Completeness*, *Accuracy*, *Conformance to Expectations*, *Logical Consistency and Coherence*, *Accessibility*, *Timeliness* and

✉ Marco Rospocher
marco.rospocher@univr.it

Matteo Lorenzini
m.lorenzini@fbk.eu

Sara Tonelli
satonelli@fbk.eu

[1] Fondazione Bruno Kessler, Via Sommarive 18, Trento, Italy

[2] University of Trento, Via Sommarive 9, Trento, Italy

[3] Università degli studi di Verona, Lungadige Porta Vittoria, 41, Verona, Italy

Published online: 23 April 2021

Table 1 Example of high-quality and low-quality descriptions from the dataset we built starting from Cultura Italia portal

Quality	Record ID	Original Italian Description	English Translation
High	iccd2225343	Dipinto entro cornice lignea verniciata ocra con bordo interno dorato. Amedeo III è raffigurato di profilo in armatura scura con ceselli in oro, mascheroni dorati sulle spalle e sull'elmo, cimiero con piume rosse e bianche. Nella parte inferiore del dipinto fascia con iscrizione a caratteri stampatello. Personaggi: Amedeo III di Savoia	Painting within an ocher painted wooden frame with a inner golden border. Amedeo III is depicted in profile with a dark armor chiseled in gold, golden figurehead on the shoulders and on the helmet. Crest with white and red plumage. On the lower part of the painting inscription with block letters. Characters: Amedeo the 3rd of Savoy
Low	work82865	Congdon si è raramente dedicato al disegno come forma espressiva autonoma, così la mole di disegni raccolti sui taccuini non sono altro che appunti visivi presi durante numerosi viaggi. In questo senso non è possibile, se non raramente, assegnare al singolo disegno un'opera finita direttamente corrispondente, così questi disegni non vengono nemmeno ad essere schizzi preparatori. La sommatoria di tutti i disegni relativi a un luogo danno origine a una serie di dipinti che non hanno un corrispettivo oggettivo nei disegni stessi. Tutto questo giustifica la presenza degli appunti all'interno delle immagini (colori, sfumature e spiegazioni di vario genere). Nel caso probabile veduta di Napoli eseguita durante un viaggio del 1951.	Congdon has rarely devoted himself to drawing as an autonomous expressive form, so the drawings in his notebooks are nothing more than visual sketch taken during his numerous trips. Rarely it is possible to assign to the single drawing the corresponding attributes as finished art work since they represents the base idea for others drawings or paintings. The collection of all the drawings related to a place give rise to a series of paintings that do not have a direct mapping to the drawings themselves. All this justifies the presence of notes inside the images (colors, shades and explanations of various kinds). In this case, probably, a view of Naples from 1951

Provenance. The work presented in this paper addresses the evaluation of the *Accuracy* dimension, defined as follows by Bruce and Hillmann: *'the metadata should be accurate in the way it describes objects. The information provided in the value needs to be correct and factual'*. In general terms, metadata accuracy is measured as the extent to which the data values in the metadata record match with the characteristics of the described object [36]. In this work, we focus in particular on determining the accuracy of the textual description (typically encoded using the *dc:description* element from the Dublin Core[1] metadata schema) of a given cultural heritage object. More specifically, we propose to assess the accuracy of such description metadata by determining whether the field contains a high-quality or low-quality description of the considered object, measured as the compliance of the textual content with the description rules from *Istituto Centrale per il Catalogo e la Documentazione* (ICCD), adopted in the Cultura Italia portal.[2]

As a first step in this direction, we create a large dataset of object descriptions, which we (semi-)automatically label as being of high quality or not. An example of high-quality and another of low-quality descriptions are reported in Table 1. In the first, all and only the necessary information related to the object (e.g. the frame) and the subject (the person portrayed in the painting) is reported. The second description, instead, is a lengthy text that focuses first on the painter and only towards the end mentions the subject of the painting. More details on the methodology and guidelines we followed for judging the quality of a description are discussed in Sect. 3.

As a second contribution, we exploit natural language processing techniques and machine learning to create a binary (high-quality vs. low-quality) classification model that is able to assess the quality of unseen descriptions by predicting the class they should belong to. To this purpose, two different classification algorithms are compared—support vector machine (SVM) [8] and the FastText logistic regression classifier [17]—leveraging the representation of descriptions as word embeddings, i.e. as real-valued vectors in a predefined vector space that compactly captures meaning similarity. The comparison is performed on three different cultural heritage domains, i.e. visual artworks, archaeology and architecture. While text analysis and machine learning have already been applied to metadata quality assessment [26], recent advances in language modelling, in particular the use of word embeddings [24], have not been explored for the task. This novel way to capture the semantic content of descriptions, together with supervised machine learning, is exploited in this work with the goal to provide some insights into which techniques and algorithms can be effectively used to support curators in the manual quality control of cultural heritage descriptions. Our goal is also to provide guidance in the creation of datasets for performing this task in a supervised setting, taking into account also the characteristics of differ-

[1] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/.

[2] http://www.culturaitalia.it.

ent domains. Specifically, we address the following research questions:

- *Research Question 1 (RQ1)* Which machine learning algorithm should be used to assess the quality of cultural heritage descriptions approximating as much as possible human judgement?
- *Research Question 2 (RQ2)* Can a classification model trained with descriptions in a given cultural heritage domain be effectively applied to automatically assess description quality in other domains?
- *Research Question 3 (RQ3)* How many annotated resources are needed to create enough training data to automatically assess the quality of descriptions?

RQ1 is addressed by comparing different classification algorithms and natural language processing techniques. With RQ2 we investigate how classification performance changes when using data from different domains, even in a combined way. Finally, with RQ3 we aim to provide guidance in applying supervised techniques to novel datasets, by assessing how the dimension of the training data affects classification quality, and therefore suggesting how many instances should be manually annotated.

Methodologically, we followed the standard best practices adopted in experimental work assessing the performance of automated processing systems. First, given the lack of an adequate resource, we developed a dataset for training and testing machine learning approaches: the dataset consists of object descriptions manually labelled by an expert annotator as high/low quality according to the adherence to the cataloguing guidelines of the digital repository indexing the objects. Second, we run several experiments to address the aforementioned research questions, assessing system performances using well-know metrics (i.e. precision, recall, F1-measure) and adopting evaluation protocols aiming to reduce possible biases (i.e. cross-validation setting, removal of duplicates). Finally, we analyse the learning curve of the best classification model, by incrementally adding new instances to the training data.

The paper is structured as follows. In Sect. 2, we introduce the problem of metadata curation, discussing the state of the art concerning past attempts to computationally evaluate metadata quality in the cultural heritage domain. In Sect. 3, we describe how the datasets for the proposed classification methodology have been selected and annotated to provide a training and test set composed of more than 100K descriptions covering three domains. Sections 4 and 5 present the classifiers used to perform the quality assessment task and the experimental settings adopted, including the evaluation measures. Section 6 presents the results of our experiments and discusses the evaluation with respect to the three research questions. In Sect. 7, we discuss findings and limitations of our approach, while in Sect. 8 we present our conclusions.

2 State of the art

2.1 Metadata quality frameworks

Despite the key role played by metadata in cultural heritage collections, evaluating their quality and establishing measures able to identify the data features that need to be improved is still a debated argument. Day [10] assessed metadata quality in e-print archives according to functional requirements defined at two separate levels: compliance with the specifications of the metadata schema used to describe the digital objects and compliance with the needs of the end user. At the first level, an object must be described strictly following the rules and guidelines of the metadata schema (or application profile) in order to be considered correct. The second, higher level of correctness requires the rightness of the values of the metadata fields: e.g. the Italian painting *The birth of Venus* by Sandro Botticelli is also know as *The Venus*. According to the second level of evaluation, both titles should be considered appropriate even if the only correct one is *The birth of Venus*. Hence, according to this second level, *quality and correctness are about fitness for purpose*.

Another approach to assess metadata quality has been defined by the NISO Foundation[3] and addresses the problem in the context of metadata creation by machines and by professionals who are not familiar with cataloging, indexing or vocabulary control [30]. The NISO Framework of Guidance for Building Good Digital Collections presents six principles of what are considered "good" metadata [27]. However, these criteria and principles do not provide a clear number of well-defined quality dimensions, so that metadata curators and end users are not supported in addressing these issues.

The first attempt to operationally define what the evaluation of metadata quality is can be found in the Metadata Quality Framework developed by Bruce and Hillmann [5], where seven dimensions and related characteristics are introduced and described, namely *Completeness*, *Accuracy*, *Conformance to Expectations*, *Logical Consistency and Coherence*, *Accessibility*, *Timeliness* and *Provenance*. However, there is no formal definition about the quality aspects that should be measured by each dimension. The authors note that it is not possible to state which of the seven dimensions they describe is most important for a given application, since the importance of each quality criterion is strictly influenced by the nature of the resource to be described, as well as by the environment in which the metadata is to be constructed or derived. Thus, great emphasis is put on the fact that percep-

[3] https://www.niso.org/standards/.

tion of quality strictly depends on context. As a consequence, metadata curators are required to follow a generic and "fitness for use" workflow [3] based on personal interpretation and manual intervention: they should check the content of each record and, depending on the types of issues, report errors to metadata creators or fix the metadata themselves, relying for instance on a controlled vocabulary. Given the growing amount of digital cultural heritage records available, this is a very time-consuming process, which cannot be adopted at scale. Concerning *accuracy*, which is the central topic of this paper, Bruce and Hillmann's framework points to the fact that "The information provided about the resource in the metadata instance should be as correct as possible [...] Typographical errors, as well as factual errors, affect this quality dimension." This is however a very narrow definition of accuracy, which only takes into account some surface features of a description (e.g. presence of mistakes), without considering that a description can be formally perfect without containing useful information, therefore being of low quality.

Besides the framework by Bruce and Hillmann, few other approaches have been proposed to automatically compute quality metrics. The ones that are more related to our work are the Framework for Information Quality Assessment by Stivlia [36], the Metadata Quality Framework by Ochoa and Duvall [28] and the Metdatata Quality Assurance Framework by Péter Király [18,19]. Other frameworks (e.g. [25]) do not include accuracy and are therefore not discussed in this paper.

Stvilia proposes a framework which overlaps with the Metadata Quality Framework by Bruce and Hillmann. The author identifies four major sources of information quality problems: *mapping, changes to the information entity, changes to the underlying entity or condition, and context changes*. To address mapping, Stvilia adopts the definition from Wand [39] according to which mapping issues arise when there is incomplete or ambiguous mapping between the information source and the information entity from the metadata schema. Changes, instead, may occur in the information entity itself or in the real-world entity it represents. Based on that, the authors develop a taxonomy of 22 dimensions, systematically organized into three categories: *intrinsic* i.e. dimensions that can be assessed by measuring information aspects in relation to reference standards (e.g. spelling mistakes); *relational*, i.e. dimensions that measure relationships between the information and some aspects of its usage (e.g. accuracy); *reputational*, i.e. dimensions that measure the position of an information entity in a given structure (e.g. authority). However, there is no implementation of these dimensions as algorithms that can be operationally applied to different cases.

Ochoa and Duvall's framework is inspired by the parameters introduced by Bruce and Hillmann and Stivlia. However, it is more detailed and specific, in that it presents several automatic calculable metrics of quality associated with the seven parameters in Bruce and Hillmann's framework. The authors point out that the proposed metrics are not intended to be a comprehensive or definite set, but should be considered as a first step towards the automatic evaluation of metadata quality.

Regarding accuracy, Ochoa and Duvall define it as *'the degree to which metadata values are "correct", i.e. how well they describe the object.'* [28]. Similar to our approach, they make use of text processing techniques and apply them to textual fields of metadata. However, they propose a general unsupervised method based on vector space model (VSM), aimed at finding the semantic distance between two resources according to the keywords stored in a vocabulary, while our approach is supervised and does not rely on external resources, because this information is already inferred by the trained classification model. Furthermore, Ochoa and Duvall's proposal to assess metadata accuracy may be affected by issues related to the length of the descriptions. Longer text contains more words than shorter ones, and this has an impact on the computation of the semantic distance with the keywords stored in the external vocabulary: the longer the text, the higher the chances that it contains some of the keywords in the vocabulary, and thus, the higher the accuracy score (due to the way the VSM works), independently from whether such keywords accurately describe the content of the text. This way, lengthy (but not accurate) descriptions containing many keywords may score higher accuracy than shorter (but accurate) descriptions. Moreover, Ochoa and Duval present also three validation studies to evaluate the proposed metrics with respect to human-made quality assessment. In general, the quality metrics do not seem to correlate with human ratings.

The third metadata quality framework we consider has been developed in collaboration with the Data Quality Committee (DQC) from the European Digital Library "Europeana"[4] by Péter Király [19]. The Metadata Quality Assurance Framework is an ongoing project tailored to measure the metadata quality of the Europeana digital library and based on Europeana Data Model (EDM) metadata profile. The framework consists of 4 different metrics, namely *completeness*, *multilinguality*, *uniqueness*, i.e. frequency of the duplicated values and *record patterns*, i.e. density distribution of filled fields among all Europeana content providers. While a lot of emphasis is put on the issue of multilinguality, which has become very relevant in data aggregation projects like Europeana, the issue of accuracy is not introduced as a separate metric, but only mentioned as a dimension that can be inferred from the others. In this respect, our work adopts a different perspective on the issue.

[4] https://www.europeana.eu/portal/en.

2.2 NLP and machine learning for description quality

We are interested in automatically assessing the quality of descriptions in digital records. The topic has already been tackled in the past with the use of machine learning and NLP, but using techniques that are different from what we propose. In [9], for example, description length is considered as a proxy for accurate content description and is used as a feature in a supervised classification task. No semantic information is analysed. In [13], string matching is used to detect information redundancy in metadata collections, a task related to metadata quality because redundancy may hinder basic digital library functions. In [33], accuracy is computed on public government data as the distance between the format of the referenced resource and the actual data type. Again, this measure is based on a formal check, without looking at what information is actually presented in the description. In [23], instead, the authors highlight the importance of a semantic check for quality assessment and therefore propose to verify correctness, completeness and relevance of metadata by creating logic rules to model relations among digital resources. In [26], the authors show that enriching the subject field with automatically extracted terms using topic modelling is valuable, especially when coupled with a manual revision by human curators. None of the techniques considered in our work, in particular supervised classification using word embeddings, has been applied and tested for Cultural Heritage repositories and resources.

3 Dataset description

In order to train a supervised system to assess metadata quality, a large set of example data is needed. Such data must be representative of the domain of interest and be manually labelled as high quality or low quality. Our use case focuses on the Italian digital library "Cultura Italia",[5] which represents the Italian aggregator[6] of the European digital library Europeana. It consists of around 4,000,000 records including images, audio visual content and textual resources. The repository is accessible via the OAI-PMH handler[7] or via the SPARQL[8] endpoint. By using the textual description encoded by the *dc:description* element from the Dublin Core metadata schema, we collect a dataset of 100,821 descriptions, after duplicate removal. These records include mainly data from "Musei d'Italia" and "Regione Marche" datasets, which have been chosen because they contain a high number of non-empty *dc:description* elements.[9] Duplicates were removed for two reasons: this reduced annotation effort in the subsequent manual annotation, and avoided that the same example appear both in the training and in the test set, a situation that could make classification biased and lead to inaccurate evaluation in supervised settings.[10] Duplicated descriptions were mainly short and of low-quality, reporting few generic words to describe an item (e.g. "Mensola.", "Dipinto.").

All these descriptions are about objects of different typologies and from different domains, a piece of information which is encoded by additional PICO[11] metadata, a qualified Dublin Core specification consisting of 91 elements.[12] Thus, leveraging the additional PICO metadata, we further organize the descriptions in three specific domains: Visual Art works (VAW) (59,991 descriptions), Archaeology (Ar) (29,878 descriptions) and Architecture (A) (10,952 descriptions).

To determine the quality of the collected descriptions, we rely on the standard cataloguing guidelines provided by the Istituto Centrale per il Catalogo e la Documentazione[13] (ICCD), i.e. the same guidelines that should be followed by the data providers of Cultura Italia portal. More precisely, a specific section of the guidelines[14] addresses how to describe any cultural item, clarifying that both the object and the subject of the item must be presented in the description as follows:

Object : the object typology and shape must be described. To describe the object, the cataloguer must refer to the vocabularies provided by ICCD, using specific terminology (e.g. the technique used for paintings and drawings, or the material for the archaeological items);

Subject : the cataloguer must report the iconographic and decorative settings of the item, such as the characters of the depicted scene in a painting and their attribution. Other aspects (e.g. the history behind the painting or the painter) should not be included.

[5] dati.culturaitalia.it

[6] Cultura Italia, in turn, is a content provider of the European digital library "Europeana".

[7] http://dati.culturaitalia.it/oaiprovider?language=it.

[8] http://dati.culturaitalia.it/sparql.

[9] Only 47.8% of the resources of Cultura Italia have a filled dc:description element.

[10] This is a technical aspect to address in order to properly assess the classification performance, and does not hinder the application of the approach for assessing the quality of descriptions in collections where multiple items share the same textual content.

[11] http://purl.org/pico/1.1/picotype.xsd.

[12] http://www.culturaitalia.it/pico/thesaurus/4.3/thesaurus_4.3.0.skos.xml.

[13] http://www.iccd.beniculturali.it.

[14] OA card, DESO and DESS element: http://bit.ly/ICCD_OA_card.

Table 2 Number of descriptions per domain labelled as high quality or low quality. Low-quality descriptions have been identified both manually and following an automatic selection

Dataset	High-quality	Low-quality (manual)	Low-quality (auto)	Total
Visual Art Work	30,383	19,824	9,784	59,991
Archaeology	19,280	6,334	4,264	29,878
Architecture	6,908	1842	2,202	10,952
Overall dataset	56,571	28,000	16,250	100,821

Following the above cataloguing guidelines, each textual description in our dataset is (semi-)automatically annotated as "High Quality" if object and subject of the item are both described according to the ICCD guidelines, and as "low quality" in all other cases. Other criteria for determining the quality of a textual description may be adopted, related for instance to the grammatical, lexical and semantic aspects of the text. In line with the working *accuracy* definition for textual metadata by Ochoa and Duval [28], in our work we focus on the compliance of descriptions with the ICCD guidelines, as discussed in Sect. 2. The annotation is carried out by an expert in cultural heritage who collaborated in the past with Cultura Italia and has therefore in-depth knowledge of the data characteristics and of the ICCD guidelines.

For each harvested description, the annotator performs the following steps:

- If the length of the description is less than 3 words, it is labelled as "low quality" (e.g. "Painting", "Rectangular table", "View of harbour"). This is done automatically based on the assumption that in few tokens it is not possible to describe both the object and the subject of a record. This concerns 5,349 descriptions, automatically labelled as "low quality";
- If there are descriptions coming from a collection not updated after 2012, they are very likely to be "low quality". This assumption is based on the annotator's domain knowledge, being aware of the history of Cultura Italia collections and therefore being able to identify less curated batches of records. This assumption is practically confirmed randomly sampling 500 records from such collections and manually checking each of them, confirming that none of the samples can be classified as "high quality". This way 10,901 descriptions are automatically labelled as "low quality";
- The remaining descriptions are then manually annotated one by one and labelled as "high quality" or "low quality".

Following best practices in linguistic annotation and dataset creation [32], we compute inter-annotator agreement, in order to assess whether the task is sound or the concept of low and high-quality metadata is too subjective. Therefore, a balanced sample of 1,500 descriptions from the dataset was sent to the metadata curator team of Cultura Italia, to be manually annotated also by one of their members. We then compared our annotation with the one from Cultura Italia. The inter-annotator agreement, computed according to Cohen's kappa [20], shows a very high level of agreement (16 diverging annotations over 1,500 description, $\kappa = 0.979$) between the two annotators. This confirms that the task can be confidently carried out by domain experts and that the quality of the resulting annotations is accurate.

Table 2 summarizes statistics of the annotated dataset and the size of the three domains. We show in a separate column ("Low-Quality (auto)") the number of descriptions with poor quality automatically identified based on their length or the year of the last update, as described above. Although low-quality descriptions are less represented than high-quality ones, there are enough examples in both classes to train a supervised system. Regarding human effort, the manual labelling task spanned around two years (partial time), at a pace of approximately 150 annotations per hour. The resulting annotated dataset is publicly available [21] under the terms of the Creative Commons Attribution-ShareAlike 4.0 Generic (CC BY-SA 4.0) licence.

4 Classification framework

Based on the data described in Sect. 3, we aim at developing an approach that can automatically identify high-quality and low-quality descriptions in cultural heritage records. We cast the problem as a binary classification task, using the annotated data to train a supervised system able to assign an unseen description to one of the two classes (low quality vs. high quality).

Classification algorithms work with numerical features, i.e. they represent each input object as a vector of real numbers which are used to build the model and to predict the class for unseen instances. Therefore, since our input data are natural language descriptions, we first convert them into numerical vectors using the FastText word embeddings [4]: each word is assigned to a real-valued vector representation for a predefined fixed sized vocabulary, capturing the fact that words that have similar meaning have a similar vector representation, and the vector representation for each description (i.e. a collection of words) is obtained by averaging the vector representations of its words. The vector representation of

each description can then be directly fed to machine learning classification algorithms.

We experiment and compare two algorithms: support vector machines (SVM) [8] and the FastText multinomial logistic regression classifier [17] (hereafter, MLR_{ft}). Both approaches are only fed with the FastText embeddings [4] as input features. This means that no manually engineered features have been used, but only those represented through the word embeddings. We remark that in the FastText word embeddings, each word is represented as a bag of character n-grams in addition to the word itself, so that also out-of-vocabulary words (i.e. words never seen during the training of the model) are included in the representation, and information on suffixes and prefixes is captured.

Before sending the descriptions to the classifiers, a preprocessing step is performed, following best practices in text classification:

- *Stopword removal* Stopwords include all terms that do not convey a semantic meaning such as articles, prepositions, auxiliaries, etc. These are removed from each description by comparing each token against a pre-defined list of Italian words imported from the NLTK Python library.[15]
- *Punctuation removal* Following the same principle of stopword removal, each punctuation is removed from the descriptions.

All the code used for running the classifiers and preprocessing the dataset is available on the GitHub code repository of the paper.[16]

4.1 Support vector machine (SVM)

Considering a binary classification problem, SVM learns to separate an n-dimensional space with a hyperplane into two regions, each of which corresponds to a class. The idea behind SVM is to select the hyperplane that provides the best generalization capacity: the SVM algorithm first attempts to find the maximum margin between the two data categories and then determines the hyperplane that is in the middle of the maximum margin. Thus, the points nearest the decision boundary are located at the same distance from the optimal hyperplane [1,34,35]. Different kernels (i.e. learning strategies) can be used in a SVM, such as radial basis function (RBF) or linear: for our task, we determined the best kernel via grid search in the classifier optimization phase. We applied SVM using the implementation available in the scikit-learn library [29].

Since the classifier takes a feature vector in input, we convert each record description into a FastText embedding. The embedding of each description is built by averaging the FastText word embeddings of the single words in the description. For this step, we rely on pre-trained continuous word representations, which provide distributional information about words and have shown to improve the generalization of models learned on limited amount of data [7]. This information is typically derived from statistics gathered from a large unlabelled corpus of textual data like Wikipedia or the GigaWord corpus. In our case, since our descriptions are in Italian, we compare two different models, a *domain-specific* and a *general-purpose* one. The first is obtained by creating FastText embeddings from the corpus obtained by merging all textual descriptions used in our experiments, while the second is the Italian pre-trained model of FastText embeddings,[17] created from Wikipedia. Both models were trained in the same way, i.e. using continuous bag-of-word with position-weights, in dimension 300, with character n-grams of length 5, a window of size 5 and 10 negatives. We also experiment with two different vector dimensions: 300, i.e. the default FastText number of dimensions, and 50, which we obtain by applying principal component analysis (PCA) [38] to the 300-dimensional embeddings.

4.2 FastText implementation of the multinomial logistic regression (MLR_{ft})

A second classification algorithm we consider is the implementation of multinomial logistic regression included in the FastText library[18] [17]. This is a linear classifier, developed by the Facebook Research Team, that was evaluated on various classification tasks (e.g. sentiment analysis, tag prediction) achieving performance score comparable to advanced deep learning models in terms of accuracy, but orders of magnitude faster for training and evaluation.

Like in the SVM scenario, we compare two variants of MLR_{ft}: one fed with the FastText embeddings obtained by merging all textual descriptions of our corpus, and one fed with the Italian pre-trained FastText embeddings created from Wikipedia. Also in this case, embeddings of different dimensions, i.e. 300 and 50, are created and compared.

4.3 Baseline

As a baseline, we train an SVM classifier using as single feature the length of the description in tokens, computed using the TINT tool [2]. We consider this a reasonable baseline to compare with other classifiers as,intuitively, low-quality

[15] https://www.nltk.org/.

[16] https://github.com/matteoLorenzini/description_quality.

[17] https://fasttext.cc/docs/en/crawl-vectors.html.

[18] https://fasttext.cc/.

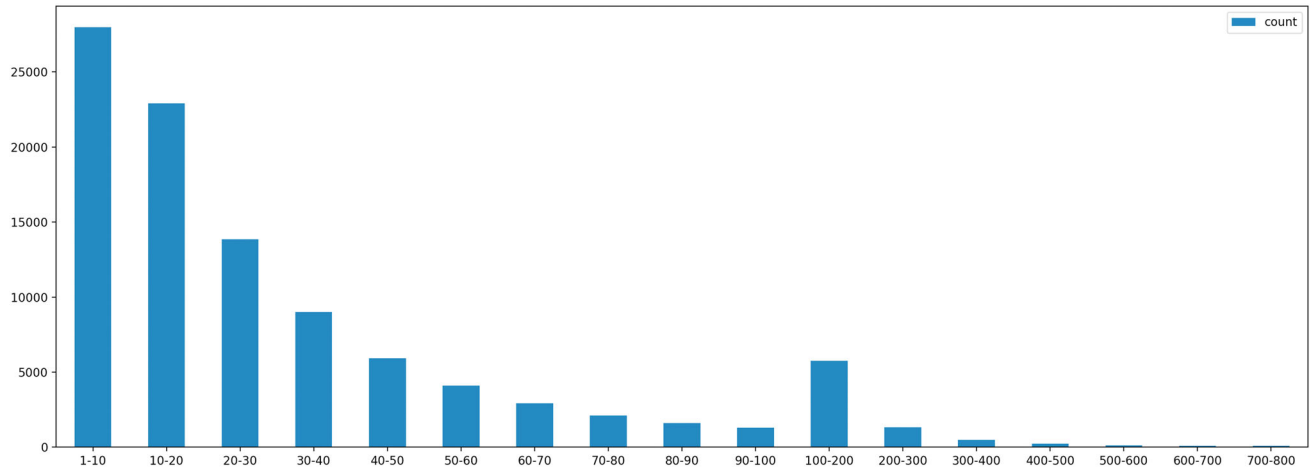

Fig. 1 Number of records in the annotated dataset (y-axis) per description length bin (x-axis) measured in tokens. Note that a bin size of 10 is used up to length 100, while a size of 100 is used for the remaining bins

descriptions tend to be shorter than accurate ones, so we want to assess whether this feature alone could be a good indicator of the description quality. In order to provide also an overview of the description length of the annotated dataset, we display a barplot in Fig. 2: on the x-axis the different length bins are reported, while on the y-axis the number of objects in the annotated dataset having the corresponding length range are shown.

5 Experimental setup

5.1 Parameter setting

We run our classification experiments on the three domains in isolation (Visual Art Works, Archaeology and Architecture) and then on the whole dataset. We compare SVM and MLR_{ft}, considering word embeddings of 50 and 300 dimensions in two variants: domain-specific and general-purpose.

All experiments are run using ten-fold cross-validation. This means that the dataset was first randomly shuffled and then split (preserving the same high-quality/low-quality proportion of the whole dataset) into 10 groups. Each group was used once as test set, while the remaining ones were merged into a training set. The evaluation scores obtained on each test set are then averaged to obtain a final, single performance evaluation.

For the SVM, three parameters need to be set, i.e. cost (C), gamma (G) and the Kernel to use. We computed them for each in-domain training set by using the grid search function in scikit-learn. The best parameter combination, which we then adopted in our experiments, is reported in Table 3. With MLR_{ft}, instead, we use the predefined hyper-parameter setup concerning *learning rate*, *epoch*, *n-grams* and *bucket*.

Table 3 SVM C, G and Kernel parameter settings used on each dataset, as result of grid search optimization

Dataset	C	G	Kernel
Visual Art Works	3	3	RBF
Archaeology	3	3	RBF
Architecture	32	8	RBF
Entire dataset	1	3	RBF

5.2 Evaluation measures

We evaluated the performance of the classifiers using a standard approach for binary tasks: we first compute Precision, Recall and F1 on each of the two classes separately (i.e. high quality and low quality) and then average them. In a 10-fold cross-validation setting, the above evaluation metrics are computed on each fold, and then averaged. More specifically, for each class we count: true positives (TP)—correctly recognized class examples; true negatives (TN)—correctly recognized examples that do not belong to the class; false positives (FP)—examples that were incorrectly assigned to the class; and false negatives (FN)—examples of the class that were not recognized. Then, *Recall*, *Precision* and *F1* are computed as follows:

- Recall $(R) = \frac{TP}{TP+FN}$. It measures how extensively a certain class is covered by the classifier;
- Precision $(P) = \frac{TP}{TP+FP}$. It measures how precise a classifier is, independently from its coverage;
- $F1 = 2 \times \frac{P \times R}{P+R}$.

Overall measures are then obtained by (macro) averaging the scores of both classes. All the metrics are computed using the Python scikit-learn "classification_report" method.[19]

6 Evaluation results

In our evaluation, we address the three research questions introduced in Sect. 1.

6.1 RQ1: Which machine learning algorithm should be used to assess the quality of cultural heritage descriptions approximating as much as possible human judgement?

We report in Table 4 the classification results obtained with the different algorithms and configurations presented in the previous sections. We include both the within-domain setting, i.e. training and test belong to the same domain (Visual Art Works, Archeology or Architecture), and the global one, considering the three datasets altogether. Overall, MLR_{ft} substantially outperforms SVM in every within-domain setting and configuration, with the former always achieving better F1 score over the latter (with improvements from 0.002 to 0.088 on the overall F1 score). Its performance is consistent for all single domains (best F1 scores ranging from .853 to .888), showing that it is robust despite the different topics mentioned in the descriptions. Also with SVM we observe a comparable performance in the three domains. While for SVM, however, feature vectors with 300 dimensions yield substantially better results, different embedding sizes do not affect much MLR_{ft} output. This means that, even limiting the computation to 50 features dimensions, and hence reducing training time, it is possible to reach good classification performances. The choice of different pre-trained embeddings does not seem to affect much the classification performance, with F1 scores that are substantially similar (with minor, negligible differences) when using in-domain or Wikipedia word embeddings.

When training and testing are performed on the whole dataset, combining descriptions from different domains, the overall scores are lower than on the single domains, suggesting that description quality is something inherent to the different cultural heritage domains, an aspect we investigate more in details with RQ2 in Sect. 6.2.

The baseline results, i.e. a classifier taking into account only description length, are different in the three domains (from .508 to .562 of F1 score). For Architecture it achieves .562 F1, meaning that in most cases longer descriptions tend to correspond to high-quality ones. This is not the case for the Visual Art Work domain, instead, where description length does not correlate with high or low quality. A possible explanation for this different behaviour may be the fact that in the domain of Architecture, or even Archaeology, descriptions of the cultural artefacts tend to be more standardised, with the same kind of structure and information, therefore description length can be a good indicator of quality. This could explain also why classification performance on the Architecture and the Archaeology datasets is higher than on the Visual Art Work data, even if the latter contains more training instances. We also observe that for the Visual Art Work domain low-quality and high-quality instances can be classified with a performance which is substantially equal, while for the other domains high-quality descriptions are recognised more accurately. This difference has two possible explanations: first, the two classes are more balanced in the VAW dataset, with roughly the same amount of instances per class. Second, classification is equally challenging on the two classes because descriptions are less standardised than in the Ar and A domains.

6.2 RQ2: Can a classification model trained with descriptions in a given cultural heritage domain be effectively applied to automatically assess description quality in other domains?

In Table 5, we report a second evaluation aimed at assessing what is the impact of the different domains on classification performance. Indeed, for the first set of experiments only descriptions from the same domain were used for training and testing (with the exception of the 'All' configuration of Table 4). In this second set of experiments, we aim at assessing to what extent quality can be associated with specific domains, and what performance can be achieved by training and testing using data from different domains. In particular, we evaluate the performance of one of best scoring classifiers of Table 4 (namely, MLR_{ft} with Wikipedia embeddings of 50 dimensions) using training data from one or more domains, and testing on one or more (possibly) different domains (i.e. not among the ones used for training). The details of the various considered combinations are reported in Table 5. All experiments are conducted preventing data overlap between train and test datasets.

The results, which should be interpreted according to the dimensions of the domain-specific datasets considered, show that using out-of-domain data greatly affects classification performance. The F1 scores are in general substantially lower than the values reported in Table 4, ranging from .371 to .831. The highest value is achieved training on VAW and testing on data from all the domains, an outcome partly justified by the substantially larger size of the VAW dataset with respect to the others. The worst classification performance is achieved using data from the Architecture dataset (A) for training, both

[19] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html.

Table 4 Classification results on Visual Art Works (VAW), Archaeology (Ar) and Architecture (A) records, and on the whole dataset. Results are reported as Precision (P), Recall (R) and F1

Dataset	System	Embeddings	Dim.	Low-quality			High-quality			Overall		
				P	R	F1	P	R	F1	P	R	F1
VAW	Baseline			.505	.446	.474	.515	.574	.543	.510	.510	.508
	SVM	Wikipedia	50	.809	.762	.785	.781	.824	.802	.795	.793	.793
	SVM	Wikipedia	300	.850	.826	.838	.835	.858	.846	.843	.842	.842
	SVM	in-domain	50	.809	.762	.785	.780	.824	.802	.794	.793	.793
	SVM	in-domain	300	.850	.826	.838	.835	.858	.846	.843	.842	.842
	MLR_{ft}	Wikipedia	50	.834	.876	.854	.873	.830	.851	.853	.853	**.853**
	MLR_{ft}	Wikipedia	300	.832	.875	.853	.872	.828	.849	.852	.852	.851
	MLR_{ft}	in-domain	50	.834	.860	.847	.859	.834	.846	.847	.847	.847
	MLR_{ft}	in-domain	300	.838	.848	.843	.850	.840	.845	.844	.844	.844
Ar	Baseline			.547	.194	.286	.673	.912	.774	.610	.553	.530
	SVM	Wikipedia	50	.814	.659	.728	.830	.918	.872	.822	.788	.800
	SVM	Wikipedia	300	.850	.752	.798	.872	.927	.899	.861	.839	.848
	SVM	in-domain	50	.815	.656	.727	.829	.918	.871	.822	.787	.799
	SVM	in-domain	300	.850	.752	.798	.872	.927	.899	.861	.839	.848
	MLR_{ft}	Wikipedia	50	.861	.848	.854	.917	.925	.921	.889	.886	**.888**
	MLR_{ft}	Wikipedia	300	.862	.843	.852	.915	.926	.920	.888	.884	.886
	MLR_{ft}	in-domain	50	.860	.844	.852	.915	.925	.920	.888	.884	.886
	MLR_{ft}	in-domain	300	.861	.845	.853	.916	.925	.920	.888	.885	.886
A	Baseline			.530	.288	.373	.671	.850	.750	.600	.569	.562
	SVM	Wikipedia	50	.796	.786	.791	.875	.882	.879	.836	.834	.835
	SVM	Wikipedia	300	.816	.799	.807	.883	.895	.889	.850	.847	.848
	SVM	in-domain	50	.799	.791	.795	.878	.883	.880	.838	.837	.838
	SVM	in-domain	300	.816	.799	.807	.883	.895	.889	.850	.847	.848
	MLR_{ft}	Wikipedia	50	.845	.822	.833	.890	.905	.897	.868	.864	.865
	MLR_{ft}	Wikipedia	300	.843	.821	.831	.889	.903	.896	.866	.862	.864
	MLR_{ft}	in-domain	50	.843	.812	.828	.884	.905	.895	.864	.859	.861
	MLR_{ft}	in-domain	300	.844	.825	.834	.891	.904	.897	.868	.864	**.866**
All	baseline			.493	.255	.336	.577	.795	.669	.535	.525	.502
	SVM	Wikipedia	50	.755	.609	.674	.734	.845	.786	.744	.727	.730
	SVM	Wikipedia	300	.794	.693	.740	.782	.860	.819	.788	.776	.780
	SVM	in-domain	50	.757	.609	.675	.735	.847	.787	.746	.728	.731
	SVM	in-domain	300	.794	.693	.740	.782	.860	.819	.788	.776	.780
	MLR_{ft}	Wikipedia	50	.769	.738	.753	.801	.826	.813	.785	.782	**.783**
	MLR_{ft}	Wikipedia	300	.767	.740	.753	.801	.824	.812	.784	.782	**.783**
	MLR_{ft}	in-domain	50	.769	.734	.751	.798	.827	.812	.784	.781	.782
	MLR_{ft}	in-domain	300	.771	.732	.751	.798	.829	.813	.784	.781	.782

Bold indicates the higher value Overall F1 score for each dataset

when used in isolation and when added to data from other domains: when training on Ar+A and VAW+A, the scores are lower than when training on Ar and VAW alone, respectively.

Overall, our results show that description quality is something inherent to the different cultural heritage domains and does not hold in general, because it must be contextualized according to each domain specification. This, as already pointed out in Sect. 2, is one of the aspects not covered by the automatic evaluation approaches previously proposed in the literature. In general, it is still possible to achieve reasonably good results when a good amount of test data comes from the same domain used for training, as shown by the last two rows of Table 5.

6.3 RQ3: How many annotated resources are needed to create enough training data to automatically assess the quality of descriptions?

Since manual annotation is, in most cases, a time-consuming task (see Sect. 3), the goal of RQ3 is to check how many anno-

Table 5 Cross-domain evaluation: Classification results obtained using training data from one or more domains, and testing on one or more (possibly) different domains (i.e. not among the ones used for training)

Dataset Test	Train	P	R	F1
VAW	Ar	.653	.645	.640
VAW	A	.488	.498	.371
Ar	VAW	.644	.654	.617
Ar	A	.447	.488	.414
A	VAW	.551	.552	.550
A	Ar	.560	.562	.556
VAW	Ar+A	.610	.609	.609
Ar	VAW+A	.624	.635	.613
A	VAW+Ar	.573	.576	.572
VAW+Ar	A	.464	.494	.383
VAW+A	Ar	.637	.633	.627
A+Ar	VAW	.610	.617	.596
VAW+Ar+A	A	.661	.556	.495
VAW+Ar+A	Ar	.738	.741	.735
VAW+Ar+A	VAW	.833	.838	.831

tated resources are needed to create a good quality dataset to assess description quality. We address this question by analysing the learning curve of MLR_{ft}, that shows how much the performance improves as the number of training samples increases (from 0.5 to 100%) and therefore estimates when the model has learned as much as it can about the data.

To run this experiment, we proceed as follows. In order to be able to compare the different sizes of training data on the same test set, we manually split the whole dataset according to the classical 80–20 Pareto principle, keeping 20% of the whole dataset (roughly 20K samples out of 100K) for testing.[20] Data were split by preserving their balance both in terms of high-quality/low-quality descriptions as well as source domain. We then trained the MLR_{ft} classifier (Wikipedia, 50 dimensions) with increasing sizes of training instances, from 0.5% (∼400 descriptions) to 100% (∼80K descriptions), and computed the evaluation scores. Figure 2 plots the F1 scores obtained (y-axis) by varying the proportion of training data used (x-axis).

The F1 score consistently grows while adding more data to the training set. The higher score is obtained using all the available training material (F1 = .845). The curve substantially flattens out at about 35% of the training material (∼28K descriptions), and the F1 score is ∼.800 already with 10% of the training material (∼8K description). This means that, even if the full training set is ten times larger, the classifier does not improve with the same proportion (less than 5%). Therefore, in a scenario in which no training data are available, we would suggest a domain expert to manually annotate around 8–10,000 in-domain descriptions to still yield good classification results. At the annotation rate described in Sect. 3, developing a manually validated dataset of this size would required approximately 53–67 h of human effort.

[20] Note that this makes the results for this experiment not directly comparable with the values reported in Table 4, which instead are obtained following the cross-validation evaluation protocol. Indeed, the results plotted in Fig. 2 can be considered as a single split (but 80–20 instead of 90–10) of the 10 ones averaged in Table 4, and based on the actual split the score obtained may be higher or lower than the ones reported in Table 4.

7 Discussions

Although our classifier may still be improved, the obtained results are very promising, suggesting that an automated analysis of description quality is feasible and it would be possible to provide a first check of the descriptions in cultural heritage records before expert validation. Our results show also that more training data are not necessarily the best solution, especially if they are not from the same domain. On the contrary, around 8–10,000 annotated instances, possibly from the same domain of interest, are enough to achieve reasonably good classification performances. Another insight from our experiments is that FastText multinomial logistic regression classifier (MLR_{ft}) outperforms SVM for this task. Moreover, the domain of the pre-trained embeddings used for building the numerical vectors of the descriptions fed to the classifiers seems to have little impact on the performances, as both general domain embeddings (trained on Wikipedia) and in-domain ones achieve comparable scores.

In general, the advantage of our approach is that no feature engineering and no language-specific processing of the descriptions are needed, apart from stopword and punctuation removal. This means that this approach is easily applicable to descriptions in any language, provided that training data are manually annotated by a domain expert.

As regards the mistakes done by the classifiers, we manually inspect the wrongly classified instances produced by one of them (MLR_{ft}, Wikipedia, 50 dimension) and they almost exclusively (95% of them) fall in one of the following three categories:

- Error type A: *Descriptions containing Latin and/or Greek terms:* misclassifications in these cases (e.g. work_48470 and work_48471 in Table 6) may be due to the fact that these words are not frequent and therefore are not represented in a meaningful way in the embedding space;
- Error type B: *Descriptions only partially compliant with the cataloguing guidelines provided by the ICCD:* these descriptions are typically annotated as low quality in our gold standard, even if the description does not contain

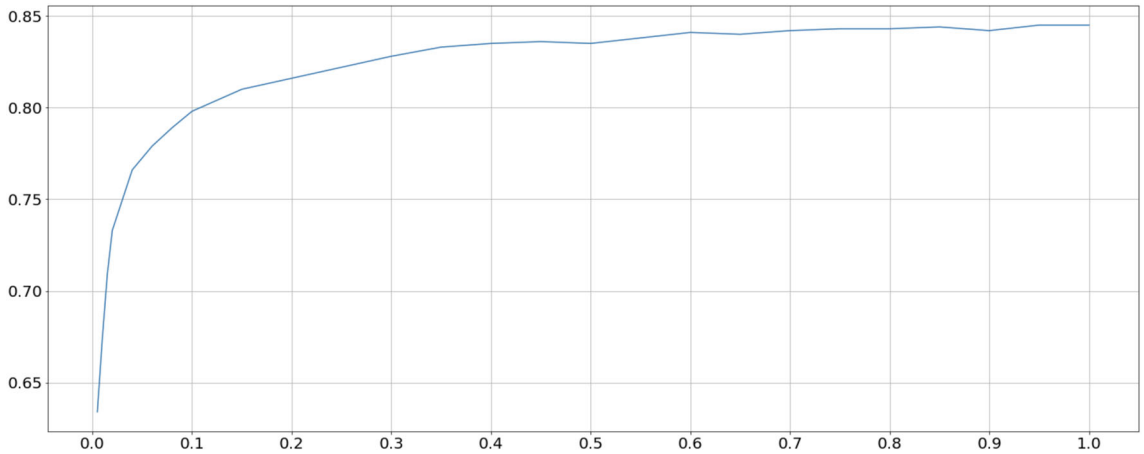

Fig. 2 Learning curve with F1 on the *y*-axis, obtained by progressively increasing the number of training instances (*x*-axis)

Table 6 Sample of high-quality (HQ) and low-quality (LQ) annotated records wrongly classified in our classification experiments

Record ID	Description	Gold	Predicted	Error
work_48470	Oinochoe *a corpo baccellato*. Applique with female protome matrix at the handle attachment.	HQ	LQ	A
124472	Black-figure painted attican Kylix , Siana type.	HQ	LQ	A
10530	Corintian Amphoriskos with zoomorphic decoration.	HQ	LQ	A
iccd3415758	The Saint, kneeled down looks up. on the bottom, to the left, there is a winged putto.	LQ	HQ	B
iccd3145858	the base lies on a parallelepiped-shaped base; [...] high volute handle.	LQ	HQ	B
iccd3165805	Brocade satin; checkered pattern. The compositional unit derives by [...] with flowers and leafs.	LQ	HQ	B
iccd3908065	Rich Oriental with mustache and half-closed mouth, head slightly oriented to [...] Figure: man	LQ	HQ	B
iccd4413810	The cycle includes three illustrated tondos, [...] .	LQ	HQ	C
iccd3913506	Wooden little angels sitting on a cloud, wrapped in a blue mantle, with wings [...]	HQ	LQ	C

Table 7 Sample of high-quality (HQ) and low-quality (LQ) annotated records correctly classified by the approach

Record ID	Description	Gold	Predicted
work_15736	The big polyptych commissioned by the Guidalotti family for their chapel [..]	LQ	LQ
work_63812	Thanks to Shearman it was verified that the painting was located in the building in via Larga where it remained [...]	LQ	LQ
iccd3906852	Crib statuette depicting an angel in a flying posture, dressed [...]	HQ	HQ
iccd2307693	[...] The man depicted has a mustache and beard and wears a wide-brimmed hat [...]	HQ	HQ

factual errors per se on the item. In our experiments, they tend to be automatically annotated as being of high quality (see for example the record iccd3908065 in Table 6);[21]

– Error type C: *Descriptions where the subject is implicit:* in these cases the classifier is not able to properly identify the domain of the item, as there may be no reference about the typology of the cultural object (see record iccd3913506 in Table 6).

Additional examples of incorrect classifications are reported in Table 6. As regards correctly classified instances, we show few examples in Table 7. Among them, the description of the Italian masterpiece "The Spring" by Sandro Botticelli (record work_63812 in the Table) consists of an

[21] The iccd3908065 description is of low quality in the gold standard according to the ICCD guidelines as it does not provide a description of the object: there is no mention in the text that the item refers to a statue, nor to its material characteristics

articulated explanation on how the painting joined the Uffizi Gallery's collection rather than describing the painting itself; hence, it has been correctly classified as having low quality by the system.

8 Conclusions

In this paper, an innovative method has been presented to automatically classify textual descriptions in cultural heritage records with the label "high quality" or "low quality". Not only we show that machine learning approaches yield good results in the task, but we also provide insights into the classifier behaviour when dealing with different domains, as well as into the amount of training data needed for classification, given that manual annotation is a time-consuming activity.

The proposed approach has several advantages: it does not require any in-depth linguistic analysis and feature engineering, since the only features given in input to the classifier are FastText word embeddings. Besides, both SVM and MLR_{ft} are less computationally intensive and energy-consuming than well-known deep learning approaches, and no specific computational infrastructure (e.g. GPU) is needed to launch the experiments. A key finding of this paper is also the importance of the domain in the classification experiments but also in the manual creation of training data: without an expert in cultural heritage, it would be impossible to create manually annotated data and to judge the performance of the classifiers from a qualitative point of view. Crowd-sourcing approaches to data annotation, which are often adopted to annotate large amounts of linguistic data through platforms such as Amazon Mechanical Turk, could not be used in our scenario, since laypeople would not have the necessary knowledge to judge the compliance of descriptions with the corresponding guidelines. This confirms the importance of multi-disciplinary work in the digital humanities, where technological skills and humanities knowledge are both necessary to achieve the project goals.

In the future, we plan to further extend this work in different research directions. As a short-term goal, we would like to compare the performance of our classifiers with other classification algorithms, including deep-learning ones. Another configuration we would like to evaluate is the use of transformer-based contextual embeddings like BERT [11] instead of word embeddings, since they provide a representation of entire chunks of text and not just at word level. This may help in better discriminating different textual contexts, i.e. dealing with different domains. An additional set of experiments could concern extending the evaluation to collections from different countries, therefore tackling descriptions in multiple languages, taking advantage of the fact that our approach does not require language-specific text processing. Moreover, another future research direction we plan to investigate is the benefit of leveraging knowledge beyond the textual content (e.g. knowledge bases, taxonomies, source authorities) to improve the assessment of description quality, especially in combination with the machine learning approaches we considered.

We see the evaluation of description quality just as one step towards a comprehensive framework for the automated assessment of metadata quality. We have already dealt with completeness in the past [22] using statistical measures. Given the promising results obtained both with *Completeness* and with description quality, we would like to operationalise other parameters proposed in the literature [5,28], again using AI-based technologies. For example, *Coherence* may be measured by cross-checking information present in different metadata fields (e.g. the content provided, when available, in the *dc:subject* field is inevitably related to the content of the *dc:description* one) using text processing and semantic web technologies.

We would like also to address a main limitation of our approach, i.e. the fact that we consider description quality as something that can be observed and measured only considering the textual component of a cultural heritage record and its compliance with ICCD guidelines. An actual assessment, with broader practical implications, should include also the item image and check the existing (or missing) correspondences between textual and visual content. This further level of analysis would require multimodal approaches, which we would like to explore as a next step in our investigation, taking advantage of existing infrastructures that support the curation of metadata, record content and images through the same interface [12].

Funding Open access funding provided by Università degli Studi di Verona within the CRUI-CARE Agreement.

References

1. Adankon, M.M., Cheriet, M.: Support Vector Machine, pp. 1303–1308. Springer US, Boston, MA (2009). https://doi.org/10.1007/978-0-387-73003-5_299

2. Aprosio, A.P., Moretti, G.: Tint 2.0: an all-inclusive suite for NLP in italian. In: Cabrio, E., Mazzei, A., Tamburini, F. (eds.) Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Torino, Italy, December 10–12, 2018, CEUR Workshop Proceedings, vol. 2253. CEUR-WS.org (2018). URL http://ceur-ws.org/Vol-2253/paper58.pdf
3. Bizer, C., Cyganiak, R.: Quality-driven information filtering using the wiqa policy framework. Web Semant. **7**(1), 1–10 (2009). https://doi.org/10.1016/j.websem.2008.02.005
4. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Trans. Assoc. Comput. Linguist. **5**, 135–146 (2017). https://www.aclweb.org/anthology/Q17-1010/
5. Bruce, T.R., Hillmann, D.I.: The continuum of metadata quality: defining, expressing, exploiting. ALA editions (2004). https://ecommons.cornell.edu/handle/1813/7895
6. Chan, L.M., Zeng, M.L.: Metadata interoperability and standardization-a study of methodology part i. D-Lib Mag. **12**(6), 1082–9873 (2006). https://dlib.org/dlib/juneo6/chqn/06chqn.html
7. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. J. Mach. Learn. Res. **12**(null), 2493-2537 (2011). https://www.jmlr.org/papers/volume12/collobert11a/collobert11a.pdf
8. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995). https://doi.org/10.1023/A:1022627411411
9. Custard, M., Sumner, T.: Using machine learning to support quality judgments. D Lib Mag. **11** (2005). URL https://dlib.org/dlib/october05/custard/10custard.html
10. Day, M., Guy, M., Powell, A.: Improving the quality of metadata in eprint archives. Ariadne **38** (2004). http://www.ariadne.ac.uk/issue/38/guy/
11. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 4171–4186 (2019). https://www.aclweb.org/anthology/N19-1423.pdf
12. Dragoni, M., Tonelli, S., Moretti, G.: A knowledge management architecture for digital cultural heritage. J. Comput. Cult. Heritage (JOCCH) **10**(3), 1–18 (2017)
13. Foulonneau, M.: Information redundancy across metadata collections. Inf. Process. Manag. **43**(3), 740–751 (2007). https://doi.org/10.1016/j.ipm.2006.06.004. http://www.sciencedirect.com/science/article/pii/S030645730600094X. Special Issue on Heterogeneous and Distributed IR
14. Gavrilis, D., Makri, D.N., Papachristopoulos, L., Angelis, S., Kravvaritis, K., Papatheodorou, C., Constantopoulos, P.: Measuring quality in metadata repositories. In: International Conference on Theory and Practice of Digital Libraries, pp. 56–67. Springer (2015). https://link.springer.com/chapter/10.1007/978-3-319-24592-8_5
15. Ishida, Y., Shimizu, T., Yoshikawa, M.: An analysis and comparison of keyword recommendation methods for scientific data. Int. J. Digital Libraries 1–21 (2020). https://link.springer.com/article/10.1007/s00799-020-00279-3
16. Jackson, A.S., Han, M.J., Groetsch, K., Mustafoff, M., Cole, T.W.: Dublin core metadata harvested through oai-pmh. J. Library Metadata **8**(1), 5–21 (2008). https://www.tandfonline.com/doi/abs/10.1300/J517v08n01_02
17. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pp. 427–431. Association for Computational Linguistics, Valencia, Spain (2017). https://www.aclweb.org/anthology/E17-2068
18. Király, P.: A metadata quality assurance framework (2015). http://pkiraly.github.io/metadata-quality-project-plan.pdf. (Research project plan)
19. Király, P., Büchler, M.: Measuring completeness as metadata quality metric in Europeana. In: 2018 IEEE International Conference on Big Data (Big Data), pp. 2711–2720 (2018). https://doi.org/10.1109/BigData.2018.8622487. https://ieeexplore.ieee.org/abstract/document/8622487
20. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**(1), 159–174 (1977). https://www.jstor.org/stable/2529310
21. Lorenzini, M., Rospocher, M., Tonelli, S.: Annotated dataset to assess the accuracy of the textual description of cultural heritage records. https://doi.org/10.6084/m9.figshare.13359104
22. Lorenzini, M., Rospocher, M., Tonelli, S.: Computer assisted curation of digital cultural heritage repositories. In: Digital Humanities Conference 2019 (DH2019) (2019). https://dev.clariah.nl/files/dh2019/boa/0807.html
23. Margaritopoulos, T., Margaritopoulos, M., Mavridis, I., Manitsaris, A.: A conceptual framework for metadata quality assessment. In: International Conference on Dublin Core and Metadata Applications **0**(0), 104–113 (2008). URL https://dcpapers.dublincore.org/pubs/article/view/923
24. Mikolov, T., Yih, W.t., Zweig, G.: Linguistic regularities in continuous space word representations. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 746–751. Association for Computational Linguistics, Atlanta, Georgia (2013). https://www.aclweb.org/anthology/N13-1090
25. Neumaier, S., Umbrich, J., Polleres, A.: Automated quality assessment of metadata across open data portals. J. Data Inf. Quality (2016). https://doi.org/10.1145/2964909
26. Newman, D., Hagedorn, K., Chemudugunta, C., Smyth, P.: Subject metadata enrichment using statistical topic models. In: Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '07, pp. 366-375. Association for Computing Machinery, New York, NY, USA (2007). https://doi.org/10.1145/1255175.1255248
27. NISO Framework Working Group (with support from the Institute of Museum and Library Services): A framework of guidance for building good digital collections. Baltimore, MD: National Information Standards Organization (NISO) (2007). URL https://www.niso.org/sites/default/files/2017-08/framework3.pdf
28. Ochoa, X., Duval, E.: Automatic evaluation of metadata quality in digital repositories. Int. J. Digit. Libraries **10**(2–3), 67–91 (2009). https://doi.org/10.1007/s00799-009-0054-4
29. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011). https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf
30. Pelaez, A.R., Alarcon, P.P.: Metadata quality assessment metrics into ocw repositories. In: Proceedings of the 2017 9th International Conference on Education Technology and Computers, pp. 253–257 (2017). https://dl.acm.org/doi/10.1145/3175536.3175579
31. Pennock, M.: Digital curation: a life-cycle approach to managing and preserving usable digital information. Library Arch. **1**, 34–45 (2007). https://www.ukoln.ac.uk/ukoln/staff/m.pennock/publications/docs/lib-arch_curation.pdf
32. Pustejovsky, J., Stubbs, A.: Natural Language annotation for machine learning—a guide to corpus-building for applications. O'Reilly (2012). http://www.oreilly.de/catalog/9781449306663/index.html
33. Reiche, K.J., Höfig, E.: Implementation of metadata quality metrics and application on public government data. In: 2013 IEEE 37th

Annual Computer Software and Applications Conference Workshops, pp. 236–241 (2013). https://ieeexplore.ieee.org/document/6605795
34. Schölkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT press (2001). https://mitpress.mit.edu/books/learning-kernels
35. Schölkopf, B., Smola, A.J., Williamson, R.C., Bartlett, P.L.: New support vector algorithms. Neural Comput. **12**(5), 1207–1245 (2000). https://www.mitpressjournals.org/doi/abs/10.1162/089976600300015565?journalCode=neco
36. Stvilia, B., Gasser, L., Twidale, M.B., Smith, L.C.: A framework for information quality assessment. J. Am. Soc. Inf. Sci. Technol. **58**(12), 1720–1733 (2007). URL https://myweb.fsu.edu/bstvilia/papers/stvilia_IQFramework_p.pdf
37. Tani, A., Candela, L., Castelli, D.: Dealing with metadata quality: the legacy of digital library efforts. Inf. Process. Manag. **49**(6), 1194–1205 (2013). https://www.sciencedirect.com/science/article/abs/pii/S0306457313000526
38. Tipping, M.E., Bishop, C.M.: Probabilistic principal component analysis. J. R. Stat. Soc. Ser. B (Statistical Methodology) **61**(3), 611–622 (1999). http://www.jstor.org/stable/2680726
39. Wand, Y., Wang, R.Y.: Anchoring data quality dimensions in ontological foundations. Commun. ACM **39**(11), 86–95 (1996). https://doi.org/10.1145/240455.240479