MACHINE LEARNING METHODOLOGY FOR AUTOMATIC METADATA ASSIGNMENT IN CULTURAL HERITAGE ARCHIVES

José Luís Machado Rei

Doctoral Dissertation
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Supervisor: Prof. Dunja Mladenić, Jožef Stefan Institute, Jamova 39, Ljubljana, Slovenia

Evaluation Board:
Prof. Tomaž Erjavec, Chair, Jožef Stefan Institute, Jamova 39, Ljubljana, Slovenia
Dr. Irena Nančovska-Šerbec, Member, Faculty of Education of the University of Ljubljana, Kardeljeva 16, Ljubljana, Slovenia
Prof. Raphaël Troncy, Member, EURECOM Graduate school and Research Centre, 450 Route des Chappes, CS 50193 - 06904 Biot Sophia Antipolis cedex, France

MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA
JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL

José Luís Machado Rei

MACHINE LEARNING METHODOLOGY FOR AUTOMATIC METADATA ASSIGNMENT IN CULTURAL HERITAGE ARCHIVES

Doctoral Dissertation

METODOLOGIJA STROJNEGA UČENJA ZA AVTOMATSKO DODELJEVANJE METAPODATKOV V ARHIVIH KULTURNE DEDIŠČINE

Doktorska disertacija

Supervisor: Prof. Dunja Mladenić

Ljubljana, Slovenia, August 2023

Acknowledgments

I would like to thank my advisor Dunja Mladenić and all my co-authors for making this work possible. A special thanks to João Pita Costa for proofreading Section 1.1, which had to be done between 21h and 23h on a weekday. Thanks also to Gregor Leban for fixing the Slovenian translation of the abstract. This research was supported in part by the European Union's Horizon 2020 research and innovation programme under grant agreements SILKNOW No 769504, Odeuropa No 101004469, and Midas No 727721.

Abstract

This thesis introduces a novel machine learning methodology for automatically assigning metadata to digitized artifacts in cultural heritage. Cultural heritage is an example of a domain that requires expert labeling, with few pre-existing labeled datasets and where simply getting more labeled data is challenging. The societal importance of cultural heritage lies in its role of safeguarding history, enhancing comprehension of the past, and nurturing a sense of belonging for current and future generations. Digitization, in turn, helps protect cultural artifacts from degradation and loss, while improving the accessibility of valuable artifacts and documents. Metadata plays a crucial role in empowering the search and exploration of large catalogs of artifacts, especially across country and language boundaries. We create a method for deducing metadata from textual descriptions of digitized items, enhancing its versatility by incorporating multiple languages and tasks to maximize the utility of existing labeled data sources using transfer learning. We augment this method with multimodal late fusion, combining text, image, and tabular classification. We show that the multimodal model significantly outperforms any single modality in metadata inference. Finally, we extend this work to a particular type of cultural heritage, literature. Literature offers a platform for exploring human emotions, thoughts, and societal issues, allowing readers to engage with different perspectives and narratives. In literature analysis, the text itself can be considered the artifact, in contrast to the analysis of text descriptions of digitized artifacts. We use semi-supervised methods to enable fine-grained multi-label emotion analysis of literature without requiring pre-existing labeled data. Emotion detection from text provides an excellent use case since emotions play a crucial role in shaping the narrative and identifying emotions in text is challenging to both humans and automated algorithms alike. We show our approach is a viable alternative to expensive and time-consuming manual labeling for creating supervised datasets. Our approach is underpinned by neural network-based representation learning, employing transfer learning in one instance and semi-supervised techniques in another. We validate our methodology through empirical findings across domains such as digitized silk fabrics, literature, and healthcare. We showcase the effectiveness of transfer learning in these diverse contexts, and underscore the benefits of multimodal approaches.

Povzetek

V tej doktorski disertaciji predstavljamo novo metodologijo strojnega učenja za avtomatsko dodeljevanje metapodatkov digitaliziranim predmetom kulturne dediščine. Kulturna dediščina je primer področja, ki zahteva strokovno označevanje, z malo že obstoječih označenih zbirk podatkov in kjer je pridobivanje dodatnih označenih podatkov izziv. Družbena pomembnost kulturne dediščine leži v njeni vlogi pri varovanju zgodovine, povečanju razumevanja preteklosti in spodbujanju občutka pripadnosti za sedanje in prihodnje generacije. Digitalizacija pa pomaga zaščititi kulturne predmete pred degradacijo in izgubo, medtem ko izboljšuje dostopnost dragocenih predmetov in dokumentov. Metapodatki igrajo ključno vlogo pri omogočanju iskanja in raziskovanja velikih katalogov predmetov, še posebej preko državnih in jezikovnih meja. Razvijamo metodo za določanje metapodatkov iz besedilnih opisov digitaliziranih predmetov, ki povečuje svojo vsestranskost z vključitvijo več jezikov in nalog z namenom maksimiziranja uporabnosti obstoječih označenih virov podatkov z uporabo učenja s prenosom znanja. To metodo dopolnjujemo z multimodalno pozno fuzijo, ki združuje besedilo, sliko in tabelarično klasifikacijo. V disertaciji pokažemo, da multimodalni model bistveno presega katero koli posamezno modalnost pri določanju metapodatkov. Nazadnje to delo razširimo na posebno vrsto kulturne dediščine, literaturo. Literatura ponuja platformo za raziskovanje človeških čustev, misli in družbenih vprašanj, kar bralcem omogoča, da se seznanijo z različnimi perspektivami in pripovedmi. Pri analizi literature je besedilo samo lahko obravnavano kot predmet, v nasprotju z analizo besedilnih opisov digitaliziranih predmetov. Uporabljamo polnadzorovane metode, ki omogočajo fino zrnato večoznačno analizo čustev literature, ne da bi potrebovali že obstoječe označene podatke. Odkrivanje čustev iz besedila predstavlja odličen primer uporabe, saj čustva igrajo ključno vlogo pri oblikovanju pripovedi in identifikaciji čustev v besedilu, kar je tako za ljudi kot za avtomatizirane algoritme izziv. Prikazujemo, da je naš pristop uporabna alternativa dragemu in časovno potratnemu ročnemu označevanju za ustvarjanje nadzorovanih zbirk podatkov. Naš pristop temelji na učenju zastopanja z nevronskimi mrežami, pri čemer v enem primeru uporabljamo prenos učenja, v drugem pa polnadzorovane tehnike. Našo metodologijo potrjujemo z empiričnimi ugotovitvami na področjih, kot so digitalizirana svilena tkiva, literatura in zdravstvo. Predstavljamo učinkovitost prenosa učenja v teh različnih kontekstih in poudarjamo prednosti multimodalnih pristopov.

Contents

List of Figures

Abbreviations

AI ... Artificial Intelligence
CH ... Cultural Heritage
CV ... Computer Vision
DA ... Domain Adaptation
GELU ... Gaussian Error Linear Unit
KG ... Knowledge Graph
LM ... Language Model or Language Modelling
LLM ... Large Language Model
ML ... Machine Learning
MLP ... MultiLayer Perceptron
MLM ... Masked Language Modeling
MTL ... Multi-Task Learning
NLP ... Natural Language Processing
NN ... Neural Network
ReLU ... Rectified Linear Unit
SSL ... Semi-supervised Learning

Chapter 1

Introduction

Over the past few decades, machine learning has evolved from a field of theoretical interest to a transformative force driving advancements across a broad range of industries and aspects of daily life. Meanwhile, the field of cultural heritage is transforming from a primarily physical and localized discipline to a dynamic and globally accessible domain, leveraging digital technologies to preserve, analyze, and disseminate the rich tapestry of human history and creativity across diverse cultures and societies. Machine learning algorithms power ubiquitous virtual assistants like Siri that millions of people use daily. They enhance the security and safety of people with applications of computer vision like assisted driving systems in mass-produced passenger vehicles. And, more recently, they became active collaborators in human creativity through applications that follow their users' written instructions to generate text, like ChatGPT, and images, like Midjourney. Much of this success has been in generic tasks for which very large amounts of labeled data can quickly be found online or created relatively cheaply or even automatically, or can strongly leverage general knowledge derived from unlabeled data. What remains to be answered is what to do about problems for which general knowledge is not sufficient, labeled data is rare or non-existent as well as too expensive to gather, and in some cases even unlabeled data is scarce. This work focuses on such problems. In particular, cultural heritage but with applications beyond it such as healthcare.

The purpose of this introduction is to provide the reader with the minimum required background information to understand the rest of this work. It is by no means an extensive treatment of the subjects involved. It begins by giving a brief explanation of the main terms present in the title of this work: machine learning (Section 1.1) and cultural heritage (Section 1.2), and what we mean by them in the context of the rest of this work. Following that, we will provide the reader with the context of our work in relation to broader aims and how it fits into other work being done towards those aims and specify our concrete scientific contributions towards those aims (Section 1.3).

1.1 Machine Learning

Machine learning (ML) is a subset of artificial intelligence (AI) that focuses on the development of algorithms and models that enable computer systems to learn and make predictions without being explicitly programmed with the instructions necessary to arrive at those predictions. The theoretical foundation of machine learning lies in statistical inference and optimization. Statistical inference provides the framework for making inferences about the underlying data distribution based on observed samples. By leveraging probabilistic and statistical techniques, machine learning algorithms estimate model parameters and learn patterns from the data. Another key concept in machine learning is the notion

of generalization. Machine learning algorithms aim to generalize from the observed data to make accurate predictions or decisions on unseen data. This is achieved by learning underlying patterns and relationships in the data while avoiding overfitting, which occurs when the model becomes overly complex and performs well on the training data but fails to generalize to new data. Regularization techniques and validation strategies are employed to strike a balance between model complexity and generalization performance. We define a machine learning algorithm in generic terms by:

Definition 1.1 (Machine Learning Algorithm). Given a set of n training examples $\mathcal{S} = \{(x_1, y_1), \dots (x_n, y_n)\}$ such that $x_i \in X$ and $y_i \in Y$ for all i from 1 to n, with x_i being an input and y_i being its corresponding output, a machine learning algorithm seeks to learn a predictive function $f : X \to Y$.

For convenience, we will write $mathcalS = \{(x, y) \in (X \times Y)\}$ under the assumption that any set of examples $\mathcal{S}$ is finite. More commonly, this is expressed by Equation 1.1, where $\hat{y}$ represents the output or prediction generated by the machine learning model for the input x, θ parameterizes our predictive function f and represents the learned parameters or weights of a machine learning model. The predictive function f is the mapping function that relates the input x to the output $\hat{y}$ based on the parameters θ.

$$\hat{y} = f(x; \theta) \tag{1.1}$$

Estimating the optimal values of the parameters θ, the "learning", involves an optimization process to minimize the discrepancy between predicted and true outputs. This process is defined by Equation 1.2, where $L : Y \times Y \to \mathbb{R}^{\geq 0}$, the loss function, represents the discrepancy (or error) between the predicted $\hat{y}$ and the true output y.

$$\min_{\theta} \quad L(Y, f(X; \theta)) \tag{1.2}$$

It is very common for the learning process to have its own parameters. These are called hyperparameters. In machine learning, a hyperparameter is a parameter whose value is set before the learning process begins. Unlike other parameters, hyperparameters are not learned from the data but are configured to control the learning process itself. Choosing appropriate hyperparameters can have a big impact on the performance of a machine learning model, and so, in practice, they are often selected through a systematic search process like grid search, random search, or Bayesian optimization. The process of finding the best hyperparameters for a given problem is known as hyperparameter tuning or hyperparameter optimization. The most common and simple example of a hyperparameter is the weight of a regularization term. We can modify Equation 1.2 to add a regularization term $\lambda R(\theta)$, resulting in Equation 1.3:

$$\min_{\theta} \quad L(Y, f(X; \theta)) + \lambda R(\theta) \tag{1.3}$$

In this idealized formulation, $R(\theta)$ is the regularization term that penalizes complex models to prevent overfitting and improve generalization to unseen examples. λ is a hyperparameter that controls the trade-off between fitting the training data and minimizing the complexity of the model. A common and simple example of regularization is to use the L^2 norm of the parameters:

$$R(\theta) = \|\theta\| \tag{1.4}$$

A loss function is any function that measures the difference, or loss, between a predicted output and a correct or true output. The specific form of the loss function depends on

the problem and the type of learning algorithm being used. For our purposes, the two most common loss functions are the Cross-Entropy Loss and its Categorical variant. The Cross-Entropy Loss defined in Equation 1.5 is used in binary classification problems, where y takes values of either 0 or 1 for the n examples in the dataset or any subset of n examples.

$$L(y, \hat{y}) = -\frac{1}{n}\sum_{i=1}^{n}(y_i \log \hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i) \tag{1.5}$$

The Categorical Cross-Entropy Loss, defined in Equation 1.6 is used in multi-class classification problems with k classes. It calculates the cross-entropy loss for each class and averages them.

$$L(y, \hat{y}) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} y_{ij} \log \hat{y}_{ij} \tag{1.6}$$

It is safe to assume that we use machine learning to accomplish something. We call what we are trying to accomplish with the use of ML a "task". Intuitively, we can also conceive that there can be a significant difference based on the context in which the task needs to be accomplished. For example, a binary (positive, negative) sentiment analysis task can be done in the context of movie reviews or in the context of financial market analysis. We call this context a "Domain". We now provide slightly more precise definitions of these concepts, based on Pan and Yang (2010):

Definition 1.2 (Domain). A domain $\mathcal{D} = \{X, P(X)\}$ consists of two components: an input space X and a marginal probability distribution $P(X)$.

In both movie reviews and financial market analysis, the input space might be English-language documents, with each document defined as a sequence of tokens. A token is a linguistic abstraction for a grouping of characters that forms a meaningful semantic unit. The probability of any one sequence of tokens, $P(x)$, will be different between the two. E.g., presumably the sequence of tokens corresponding to "ruined by bad acting and too many murders" will be more likely in movie reviews than in finance. A task refers to a specific problem that a learning system is designed to solve.

Definition 1.3 (Task). A machine learning algorithm is said to solve a task $\mathcal{T} = (Y, t : X \to Y)$ if it learns a function f that approximates t, from a finite set of examples, $\mathcal{S} = \{(x, y) \in (X \times Y)\}$ with t being an unknown function that represents the optimal mapping from an input space X to an output space Y. Alternatively, we write $f : X \to Y \approx t : X \to Y$.

Thus, a loss function L estimates the error of using f to approximate t.

1.1.1 Neural Networks

A neural network (NN) is a type of machine learning model loosely inspired by the structure and functioning of the human brain. It is a computational model composed of interconnected units known as neurons organized in layers. Each neuron receives input signals, processes them using a nonlinear activation function, and produces an output signal, which can serve as input to neurons in a successive layer or be used as the final output of the network. The core intuition in NNs is the successive building of progressively more abstract representations of the input data, also called features, through the layering of the neurons, illustrated by Figure 1.1 (Goodfellow et al., 2016). The first layer of a NN is the first hidden layer, however, it is common to refer to a "visible layer" or "input layer" that is

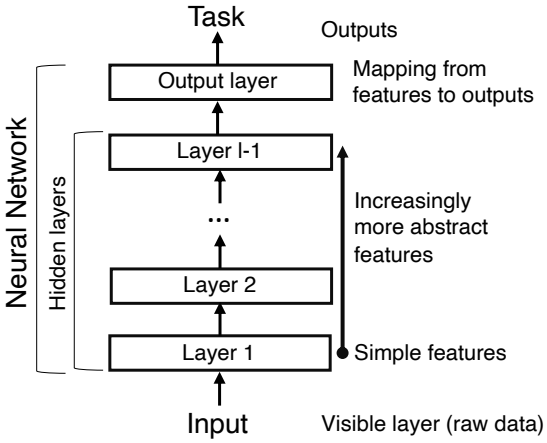


Figure 1.1: Illustration of a neural network, with each successive hidden layer progressively extracting increasingly more abstract features.

just a placeholder for the input data. Thus, the smallest NN is considered to be a "3-layer neural network" with the visible layer, a single hidden layer, and the output layer.

Mathematically, a NN is a function mapping input values to output values like any machine learning algorithm as in Definition 1.1 and Equation 1.1: $f : X \to Y$ parameterized by the weights associated with the connections between neurons, except that this function is deliberately formed by composing many simpler functions. That is, a NN with l layers can be thought of mathematically as a function f formed by the composition of l functions as expressed in Equation 1.7.

$$f(X;\theta) = f_l(\cdots f_2(f_1(X;\theta_1);\theta_2)\cdots;\theta_l) \tag{1.7}$$

This definition also gives NNs their characteristic training algorithm: backpropagation, short for "backward propagation of errors" (also called "backprop"). It performs the optimization step described in Equation 1.2 by traversing the network in reverse order, from the output to the input layer, according to the chain rule from calculus, computing only the necessary error gradients at each layer before calculating them for the next.

A layer is an abstraction and can be an arbitrarily complex function, itself composed of other functions of varying complexity. Let us consider h_i as the output of the hidden layer f_i corresponding to an input vector x. The simplest definition of a layer is the affine function followed by a nonlinear function g, called an activation function, shown in Equation 1.8:

$$f_i(h_{i-1};\theta_i = w_i, b_i) = g(w_i^\top h_{i-1} + b_i) \tag{1.8}$$

In this formulation, the parameters θ_i are composed of weights w_i and biases b_i which form the affine transformation: a linear transformation of the input via weighted sum, combined with a translation via the added bias (A. Zhang et al., 2021). The non-linearity g in the classic multilayer perceptron (MLP) is the sigmoid function σ, other common examples include the hyperbolic tangent, the Rectified Linear Unit (ReLU), and the Gaussian Error Linear Unit (GELU).

1.1.2 Learning Scenarios

ML can be subdivided into categories based on how the learning is done, also known as learning scenarios. For example, whether it is done from labeled data, unlabeled data,

or both. Labeled data refers to a dataset where each data instance (e.g., text, image, or audio) is accompanied by corresponding pre-assigned and known labels, the target output. A data instance paired with its corresponding target output forms a supervised example. Accordingly, ML can be divided into the following broad categories or paradigms (Mohri et al., 2018):

- Supervised learning involves training a model with labeled data, where the desired output is known, and the model learns to generalize from these examples to predict target outputs for new or unseen data instances.
- Unsupervised learning aims to find patterns and structures in unlabeled data without explicit target values.
- Semi-supervised learning (SSL) refers to the middle ground between the previous two categories, where explicit target values are desired but supervised examples are not sufficient, and thus it is desirable to learn from both labeled and unlabeled data.

In the context of Natural Language Processing (NLP), the landmark and widely discussed expert opinion piece entitled "The Unreasonable Effectiveness of Data" (Halevy et al., 2009) argued that the biggest successes of machine learning had been, up to then, in tasks for which massive amounts of supervised data were readily available. That same year, in Computer Vision (CV), the ImageNet database with 3.2M annotated images was announced with what became known as the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) (Russakovsky et al., 2015) first being held the following year and ultimately ushering in a revolution in CV with the publication of AlexNet (Krizhevsky et al., 2012).

Supervised learning on very large datasets remains the best option, when possible. But it is not always possible, e.g., when the data is rare or prohibitively expensive to label. Semi-Supervised machine learning addresses the problem that the labeled data may be too few to build a good model, by making use of a large amount of unlabeled data and a small amount of labeled data. Commonly, semi-supervised techniques assume that the distributions of the labeled and unlabeled data are the same. Transfer learning, in contrast, allows the domains, tasks, and distributions used in training and testing to be different (Pan & Yang, 2010). In parallel, learning just from tiny labeled data sets has also motivated other paradigms, such as Few-Shot Learning (Y. Wang et al., 2020). We adopt a slightly modified definition of transfer learning from Pan and Yang (2010) and Plested and Gedeon (2022):

Definition 1.4 (Transfer Learning). Given a learning task $\mathcal{T}_S$ in a source domain $\mathcal{D}_S$ and learning task $\mathcal{T}_T$ in a target domain $\mathcal{D}_T$, transfer learning aims to learn or improve the learning of the target predictive function $f_T(\cdot)$ on the target task $\mathcal{T}_T$ using the source task $\mathcal{T}_S$, where $\mathcal{D}_S \neq \mathcal{D}_T$ or $\mathcal{T}_S \neq \mathcal{T}_T$.

Representation learning, as described in Section 1.1.3, and multi-task learning, described in Section 1.1.4, are types of transfer learning that imply $\mathcal{T}_S \neq \mathcal{T}_T$, that can be designated as inductive transfer learning. Domain adaptation (DA), described in Section 1.1.5, and cross-lingual learning, described in Section 1.1.7, are also types of transfer learning that imply $\mathcal{D}_S \neq \mathcal{D}_T$, that can be designated as transductive transfer learning (Ruder, 2019). The theoretical difference between transductive and inductive learning lies mainly in their goals: transductive learning is about applying learned knowledge to specific instances, while inductive learning is about learning a general rule that can be applied to new unseen instances.

1.1.3 Representation Learning

Representation learning is a strategy where a model learns a representation of its input data that make it easier to extract useful information when building classifiers or other predictors (Bengio, 2012; Ericsson et al., 2022). One of the big impacts of this strategy has been in transfer learning, by learning a generally useful representation of data from one task, it is then possible to use this representation in a different task. In CV, the most common method of representation learning has been supervised representation learning on ImageNet (Oquab et al., 2014) while in NLP the most common method has been unsupervised learning based on language modeling (LM). GPT (Radford et al., 2018) learned its representation of language by predicting the next token in a text sequence, while BERT (Devlin et al., 2019) learned its representation by being trained to predict masked tokens. Both of these Large Language Models (LLMs) and subsequent models' learned representations were then evaluated by fine-tuning them on downstream tasks, such as those in the General Language Understanding Evaluation (GLUE) benchmark (A. Wang et al., 2018). Figure 1.2 illustrates the concept of representation learning: a pretraining phase, possibly unsupervised, learns a representation of input data followed by supervised fine-tuning on a target task. The target task is called "downstream" because it occurs separately and after the pretraining task. The pre-training task is also commonly called a pretext task. In NLP, it is common to call the shared part of the model the "encoder" because it encodes an input text into a real dimensional vector. In CV, it is common to call the shared model "feature extractor" because it creates feature maps from an input image. The task-specific layers for a given downstream task are commonly called the "task head" or just "head".

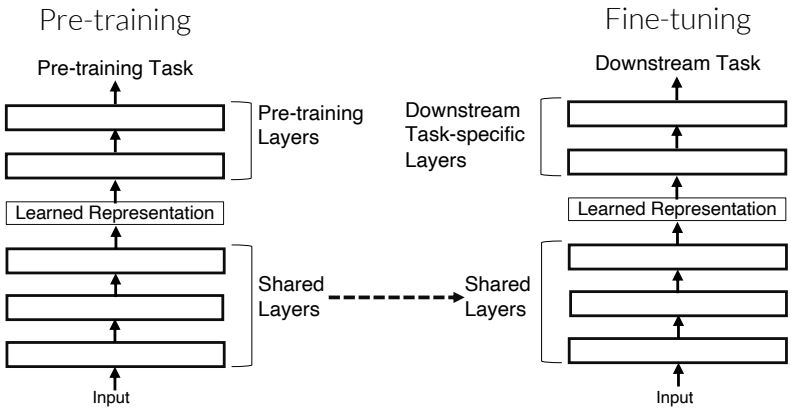


Figure 1.2: Representation Learning: Pre-training first learns a representation of the input which is later used in a downstream task.

We provide the following definition for Representation Learning:

Definition 1.5 (Representation Learning). Representation learning uses a pretext learning task $\mathcal{T}_P$ to learn a mapping function $h(x;\theta_R)$ that improves the performance of the predictive function $f(x,\theta)$ in the target learning task $\mathcal{T}_T$ when learning $f(x;\theta) = g(h(x;\theta_R),\theta_T)$.

The mapping function $h : X \to Z$ maps the input space X to a typically lower-dimensional, feature space Z, the learned representation, such that it captures important features of the input data. When considering Figure 1.2, the parameters θ_R correspond to the weights of the layers that are shared from the pre-training model to the fine-tuning

model and the parameters θ_T correspond to the weights of the task-specific layers, thus $\theta = \theta_R + \theta_T$ are the weights of the fine-tuning model (f). We can adapt Equation 1.2 to define the representation learning objective in Equation 1.9:

$$\min_{\theta_R, \theta_K} \quad L(U, k(h(X; \theta_R); \theta_K)) \tag{1.9}$$

Where $k : Z \to U$ is the predictive function that maps the learned representation Z to the pre-training target U and θ_K corresponds to the weights of the pre-training layers in Figure 1.2. Thus, $k_{\theta_K}(h_{\theta_R}(\cdot))$ can be considered our pre-training model, with k_{θ_K} often, but not always, being discarded after pre-training. In CV, when a model is pre-trained on the ImageNet ILSVRC 2012 dataset, the loss function is just the Categorical Cross-Entropy loss (Equation 1.6) and U is just the space of 1000 categories in that dataset. In NLP, the inputs to models are normally sequences of tokens and the pre-training objective is to maximize the likelihood of observed sequences of tokens. Through the use of the chain rule of conditional probabilities and by using the Markov assumption, this is equivalent to minimizing the Categorical Cross-Entropy loss for the model's prediction of the next token or a masked token in an input sequence. Thus, X are token sequences, U is the space of predefined tokens and $k_{\theta_K}(h_{\theta_R}(\cdot))$ is called a language model since it estimates the likelihood of any sequence of language tokens occurring.

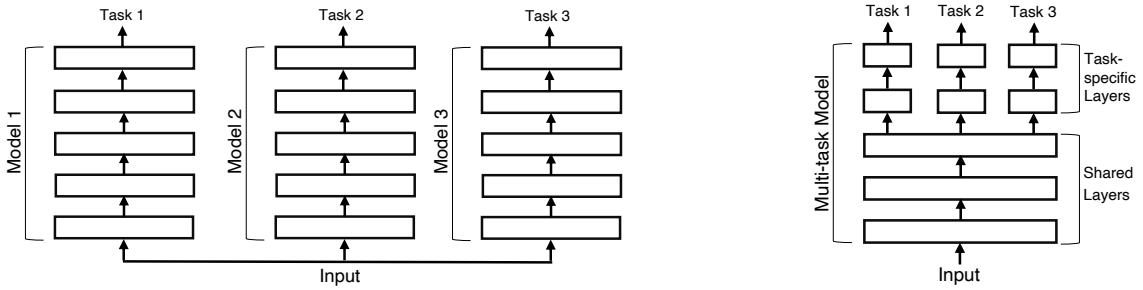
1.1.4 Multi-task Learning

Transfer is not limited to dedicated representation learning followed by target downstream tasks, but can also occur between the target tasks themselves. This is the case in Multi-task Learning (MTL) (Caruana, 1997). From a theoretical perspective, multi-task learning can be seen as a type of inductive transfer, where inductive transfer can enhance a model by introducing a bias that favors certain hypotheses over others. In MTL, the inductive bias for any task is derived from the other tasks, guiding the model to favor hypotheses that explain multiple tasks simultaneously (Ruder, 2017). As Caruana (1997) wrote: "MTL improves generalization by leveraging the domain-specific information contained in the training signals of related tasks". We adopt a slightly modified definition from Y. Zhang and Yang (2017):

Definition 1.6 (Multi-task Learning). Given m learning tasks $\{\mathcal{T}_i\}_{i=1}^m$ where each task $\mathcal{T}_i$ is defined as a mapping between an input X and output Y_i, and where all the tasks or a subset of them are related but not identical, multi-task learning aims to help improve the learning of the predictive function $f_i : X \to Y_i$ for $\mathcal{T}_i$ by using the knowledge contained in the m tasks.

Figure 1.3 illustrates single-task and multi-task learning. In neural networks, the predominant approach is hard-parameter sharing, where the first N_s layers are shared across all tasks, while the final N_t layers, including the output layer, are task-specific. Each task makes use of $N = N_s + N_t$ layers, with the model having a total of $N_s + \times \sum_{t=1}^{m} N_t$ layers. Each task can have a different number of task-specific layers, and these can also be different in size and type.

The overall MTL objective is to find the optimal model parameters that minimize the joint loss function, resulting in a model that performs well on all tasks simultaneously. Equation 1.10 which adapts Equation 1.2 to MTL. Given a set of m learning tasks, the objective is to estimate the parameters $\theta = \{\theta_S, \theta_1, ..., \theta_m\}$ where θ_S are the shared parameters and $\{\theta_1, ..., \theta_m\}$ are the task-specific parameters. This is accomplished by minimizing the joint loss L_{MTL} of the predictive function $f : X \to \{Y_1, ..., Y_m\}$ which maps a single input x to a set of outputs $\{y_1, \ldots, y_m\}$ for all m tasks:



(a) Single-task learning

(b) Multi-task learning

Figure 1.3: Single-task and multi-task learning.

$$\min_{\theta} \quad L_{\text{MTL}}(Y, f(X;\theta)) = \sum_{i=1}^{m} \lambda_i L_i(Y_i, f(X;\theta)) \tag{1.10}$$

Where L_i is the task-specific loss for task T_i, λ_i is a task-specific weight. We can also think of f as being a set of predictive functions for each task i, $f = \{f_i(X;\theta_S,\theta_i), i \in \{1, ldots, m\}\}$ with θ_S being shared among them. Another way to view this is through the lens of representation learning, where the model learns a shared representation for all tasks corresponding to θ_S, i.e, $h(x;\theta_S)$ and a series of task-specific models on top of that representation $f = \{f_i(h(x;\theta_S),\theta_i), i \in \{1, ldots, m\}\}$. Although by definition, the aim of MTL is to improve the learning of f_i for a task $\mathcal{T}_\rangle$, a multitask model may be desirable for other practical reasons, such as space and computational costs: a multitask model with shared parameters might be preferable even if for any given task it does not perform any better than a single-task learning equivalent.

MTL can of course be combined with upstream representation learning (e.g., (X. Liu et al., 2019)), the simplest case is when the shared layers are the same between the pre-training task and all downstream tasks, i.e., a shared encoder. Additionally, unsupervised representation learning can be done using MTL (Bengio, 2012). The original BERT pre-training used both the masked language modeling (MLM) task and a next sentence prediction task. In contrast, the original GPT used single-task learning for pre-training but performed MTL during fine-tuning by using both the unsupervised pre-training objective (LM) and the task-specific objective on the task data.

1.1.5 Domain Adaptation

The goal of Domain Adaptation (DA) is to leverage the knowledge learned from a source domain, where labeled data is available, to improve the performance of a model on a target domain, where labeled data may be scarce or unavailable. The source domain is typically related to the target domain, but may differ in some aspects. As discussed in Section 1.1.2, ML techniques are often categorized based on whether labeled data is available. However, in DA, data can come from either a source or target domain, which complicates that categorization. We adopt the terminology from Wilson and Cook (2020), whereby in "unsupervised" DA both labeled source data and unlabeled target data are available, and in "supervised" domain adaptation both labeled source and target data are available.

Given Definition 1.4 transfer learning for domain adaptation implies $\mathcal{D}_S \neq \mathcal{D}_T$. When labeled data from the target domain are available, the most simple and common baseline approach to domain adaptation is Mixed Training. It consists in training the model on a

mix of examples from the source domain with the few examples available from the target domain. The next improvement in Mixed training is, typically, to give more weight or oversample target domain data during the learning (Daumé III, 2007). Unsupervised transfer is necessary when no labeled data is available in the target domain. Its simplest baseline, Source-only (Daumé III, 2007), consists of transfer without adaptation, i.e., using the source predictive function $f_S(\cdot)$ as the target predictive function $f_T(\cdot)$. Several approaches have been proposed to improve upon this baseline (Ramponi & Plank, 2020; Wilson & Cook, 2020), such as reweighting examples based on similarity to unlabeled target data, data augmentation, and adversarial training.

1.1.6 Multimodal Machine Learning

Multimodal machine learning deals with the integration and analysis of data from multiple modalities or sources. Modalities can include different types of data such as text, images, audio, video, sensor data, and more. The key objective of multimodal machine learning is to leverage the complementary information present in multiple modalities to improve learning and performance in various tasks. The combination of multiple modalities is called multimodal fusion and the primary taxonomy is based on how the modalities are merged together (Baltrusaitis et al., 2019; Ramachandram & Taylor, 2017):

- Early Fusion: the information from different modalities is combined at the input level before being fed into any model.
- Late Fusion: in late fusion, the information from different modalities is handled by effectively different models and only merged at the decision level, by voting, averaging, or by a classifier that combines the decisions from multiple modalities (commonly called a metaclassifier).
- Mid Fusion: the information from different modalities is fused at a middle or intermediate stage of processing. Each modality is processed separately, and its representations or features are extracted separately. These modality-specific features are then combined or concatenated before being used as input to the subsequent layers or modules of the model. This category includes a set of significantly different approaches such as intermediate, hybrid, and cross-modality fusion.

Early fusion can leverage the complementary information from different modalities right from the start and thus can reduce the risk of information loss during feature extraction. However, it may not effectively capture higher-level interactions between modalities when they are individually complex, have different structures, and are not naturally combinable. Critically, it usually requires modalities to be aligned, i.e., an input for one modality must have a corresponding input for another (e.g., different sampling rates or missing modalities in input examples). Mid fusion techniques allow each modality to be processed individually and capture its specific patterns and characteristics before integrating the information. Each modality can contribute more independently before their combination by having modality-specific feature extraction. The majority of work in neural network multimodal fusion adopts an intermediate-fusion approach, where a shared representation layer is constructed by merging units with connections coming into this layer from multiple modality-specific paths. But as in early fusion, modality alignment remains an issue for learning algorithms, as many parameters still need to be learned for the joint part of the model. Issues can also exist with regard to dimensionality, i.e., when the dimension of the representation of a modality is much larger or much smaller than the others. Late fusion allows each modality to have its own dedicated processing, feature extraction, and decision-making mechanism. This approach allows for more flexibility in modeling and can handle varying levels of relevance and importance across modalities. Since each modality

is effectively independent and fusion can be done on lower dimensional decision space, it becomes easier to handle alignment between modalities and the dimensionality for each is the same (decision space i.e., the target outputs). The main drawback of late fusion is that it may not fully capture interactions between modalities. Figure 1.4 shows an illustration of the different approaches to multimodal fusion, with red dots indicating where fusion first occurs.

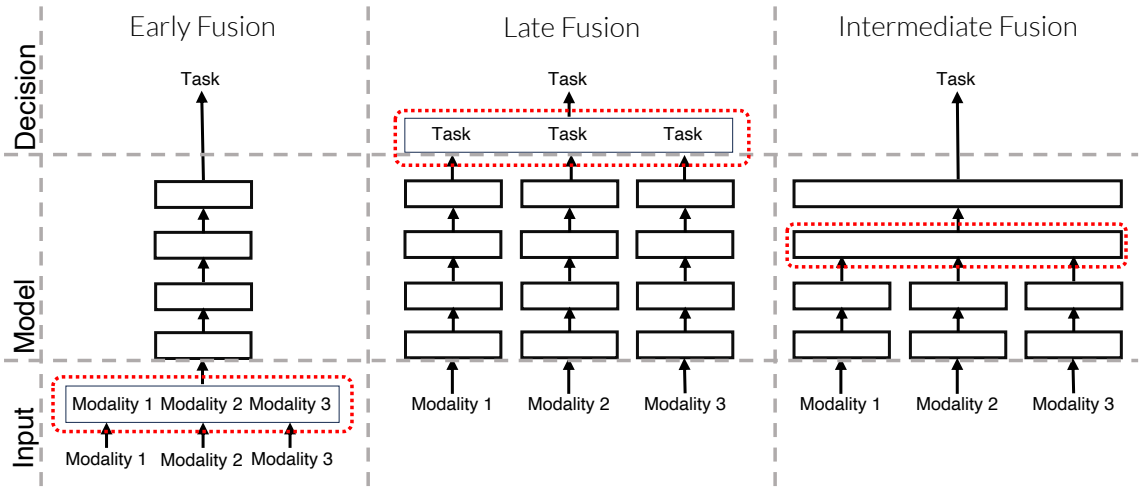


Figure 1.4: An illustration of fusion approaches for multimodal learning.

Definition 1.7 (Multimodal Learning). Given a data set $\mathcal{S}$ of size n such that $\mathcal{S} = \{(x, y) \in (X^m \times Y)\}$ where each $x \in X^m$ is an input example, each $y \in Y$ is the corresponding output, and X^m is a multidimensional input space of dimension m, i.e., $x = [x_1, x_2, \ldots, x_m]$ where each dimension corresponds to a modality, multi-task learning aims to help improve the learning of the predictive function $f : X^m \rightarrow Y$ by using the knowledge contained in the m modalities.

1.1.7 Multilingual and Cross-Lingual Models

Multilingual and cross-lingual learning are forms of transfer learning specific to machine learning fields that deal with human languages, such as NLP. Rather than a transfer between domains or tasks, the transfer occurs between languages. There exist different definitions for the terms multilingual and cross-lingual, and instances of the terms being used interchangeably (Pikuliak et al., 2021). We will provide our own definitions here. Broadly, under our definition, a multilingual model is a model that accomplishes a task in any language it has been trained to perform that task on. A cross-lingual model is a model that is capable of accomplishing a particular task with inputs in a language that it has not been trained to perform that task on (Jafari et al., 2021). This ability is often called "zero-shot" cross-linguality.

Definition 1.8 (Multilingual learning). Given l language-specific disjoint sets of training examples $\{\mathcal{S}_i\}_{i=1}^{l}$ where each set $\{\mathcal{S}_i\}$ is defined as a mapping between an input space X_i in language i and an output space Y, multilingual learning aims to improve the learning of the predictive function $f_i : X_i \rightarrow Y$ for the language i by using the knowledge contained in all l language-specific training sets.

As in multi-task learning, a multilingual model might be preferred over multiple language-specific models in practice, even if it does not improve results for any language. In this

case, the multilingual model can simply be called a "polyglot" model. The benefits can come from reduced processing or memory use from shared parameters or increased simplicity of deployment. When processing data in multiple languages, rather than having l language-specific models, only a single multilingual model is necessary. In the case of a multilingual model, $f(X;\theta)$, all θ parameters can be shared between different languages while for language-specific models, $\forall(g_i(X_i;\theta_i), g_j(X_j;\theta_j)), i \neq j \implies \theta_i \neq \theta_j$.

In cross-lingual learning, the learning, or at least most of it, can occur in one language (e.g., English), while the prediction occurs in a different language (e.g., a low resource language):

Definition 1.9 (Cross-lingual Learning). Given a set of training examples in a source language, $\mathcal{S}_i = \{(x_i, y) \in (X_i \times Y)\}$ cross-lingual learning aims to learn the target language predictive function $f_j : x_j \rightarrow Y$ where $i \neq j$.

Both multilingual and cross-lingual models rely on representation learning (Section 1.1.3). Cross-lingual models must learn a representation $h(X)$ such that $x_i \sim x_j \implies h(x_i) \sim h(x_j) \forall i, j$. That is, if any two inputs have a similar meaning, even if written in different languages, their representations should be similar in some way. This condition is more relaxed for multilingual models, though still expected to some degree. Cross-lingual learning tends to rely on some supervision for the understanding that $x_i \sim x_j$. Under our usage of the terms, multilingual learning is more about inductive transfer learning while the focus of cross-lingual is more on transductive transfer learning.

1.1.8 Vocabulary Alignment for Transfer Learning

A controlled vocabulary is a standardized set of terms used in information retrieval, cataloging, and indexing. This set of terms has been carefully chosen or curated, and its use is governed by a specific set of rules or guidelines. The goal of a controlled vocabulary is to ensure consistency and reduce ambiguity in the description and retrieval of information across various documents or datasets. Controlled vocabularies can be beneficial in fields like library science, content management, data analysis, and more. By using a consistent set of terms, researchers, librarians, or information managers can more accurately and efficiently search for and classify content. Examples of controlled vocabularies include thesauri, subject headings, taxonomies, and ontologies. A thesaurus is a tool that lists terms grouped together according to similarity of meaning and their relationships to other terms in the hierarchy, such as broader terms, narrower terms, related terms, etc. It is generally used to support indexing, tagging and searching of content. The Medical Subject Headings (MeSH)[1] thesaurus is a controlled and hierarchically-organized vocabulary produced by the United States National Library of Medicine. It is used for indexing, cataloging, and searching of biomedical and health-related information. The SILKNOW Thesaurus (Alba et al., 2022; Léon et al., 2019) was developed for organizing information related to silk textile cultural heritage, such as digital archives. Figure 1.5 shows the entry for "Silk" in the Silk Heritage Thesaurus [2].

Categorizing items into terms in a thesaurus can be defined as a learning task. In this case, the machine learning system must learn an approximation of the thesaurus function t, $f : X \rightarrow Y$ where X are items to catalog and Y are classes or terms in a thesaurus. Alignment (mapping, matching) is the process of comparing two or more thesauri to reach agreements on how to merge them. The need for reconciliation of vocabularies often arises when multiple vocabularies about the same topic disagree with each other. It is a common

[1] www.nlm.nih.gov/mesh/

[2] https://skosmos.silknow.org/thesaurus/en/

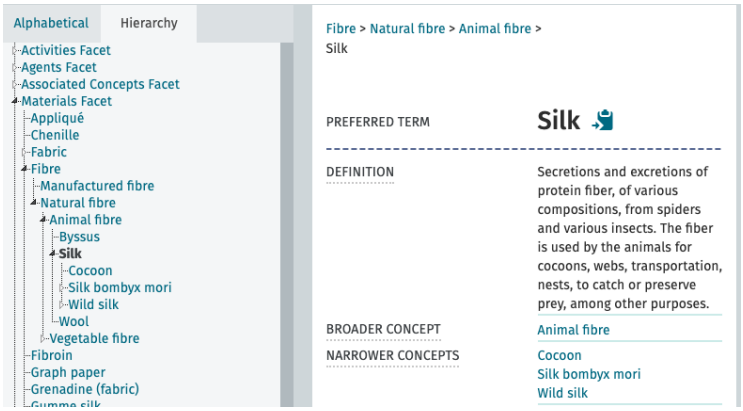


Figure 1.5: "Silk" in the Silk Heritage Thesaurus.

process in Information Science and in Semantic Web technologies, and applies to all types of controlled vocabularies such as taxonomies or ontologies (Cheng & Xia, 2023).

In the context of transfer learning, consider two controlled vocabularies a and b:

- The tasks:
  - $\mathcal{T}_a = (Y_a, t_a : X \to Y_a)$
  - $\mathcal{T}_b = (Y_b, t_b : X \to Y_b)$
- Their corresponding sets of examples:
  - $\mathcal{S}_a = \{(x, y_a) \in (X \times Y_a)\}$
  - $\mathcal{S}_b = \{(x, y_b) \in (X \times Y_b)\}$
- Their corresponding domains:
  - $\mathcal{D}_a = (X, P_a(X))$
  - $\mathcal{D}_b = (X, P_b(X))$
- Two alignment functions:
  - $g_1 : Y_a \to Y_c$ that aligns the vocabulary a with c
  - $g_2 : Y_b \to Y_c$ that aligns the vocabulary b with c

It is possible to define a task $\mathcal{T}_c = (Y_c, t_c : X \to Y_c)$ and with a slight abuse of notation, we can define two sets of examples, created by mapping $\mathcal{S}_a$ and $\mathcal{S}_b$ to c using g_1 and g_2, respectively:

- $\mathcal{S}_{a \to c} = g_1[\mathcal{S}_a] = \{(x, g_1(y_a)) \forall (x_i, y_a) \in \mathcal{S}_a\}$
- $\mathcal{S}_{b \to c} = g_2[\mathcal{S}_b] = \{(x, g_2(y_a)) \forall (x_j, y_a) \in \mathcal{S}_b\}$

In other words, we have converted $\mathcal{T}_a$ and $\mathcal{T}_b$ into the same task $\mathcal{T}_c$ by mapping their labels to a common label space. The problem of transfer learning is now strictly one of domain adaptation, since $\mathcal{T}_{a \to c} = \mathcal{T}_{b \to c}$ but $\mathcal{D}_a \neq \mathcal{D}_b$. Note that, in practice, in our context, an alignment function g may actually be a partial function. $g : Y_a \rightharpoonup Y_c$ is a function from a subset $\mathcal{S}$ of Y_a to Y_c. The case when g is not a total function, $\mathcal{S} \neq Y_a$, corresponds to some labels from the original task not being mapped to the new task. For example, if we are interested in using a large part of the MeSH vocabulary for classifying documents according to medical terms but would rather drop the subtree of terms corresponding to geographic locations. g does not need to be an injective function, which corresponds to multiple terms in the source vocabulary being mapped to the same term in the target vocabulary.

1.2 Cultural Heritage

The term application domain refers to a specific area or field of application in which ML techniques and models are deployed to solve problems or address challenges. The primary domain this work is concerned with is cultural heritage (CH). The field of CH encompasses the study, preservation, and promotion of cultural artifacts, sites, traditions, and practices that hold historical, artistic, archaeological, or societal value. It involves the exploration and understanding of various aspects of human culture, including art, monuments, literature, fashion, and more. Facilitating access to CH knowledge is the primary function of galleries, libraries, archives, and museums, collectively referred to by the acronym GLAM. These institutions collect, curate, and maintain CH for the public and researchers alike. In the last few decades, GLAMs have been transitioning towards digital tools and platforms, a process that is still ongoing. Some large or exceptionally well-funded institutions are able to carry out large digitization projects. But not all are so fortunate. Across the spectrum, most collections have not been digitized in their entirety, but a fair amount of information in various digital forms is already available. Digitization has not only made the public's access to CH easier, but crucially, it has greatly enhanced researchers' ability to conduct their work. Instead of traveling to the location of the archive and/or physically searching for the information they need inside printed catalogs, they can now access it online. Studying and documenting CH is essential for its preservation and contributes to fostering greater respect, enjoyment, and appreciation of cultural heritage within society as a whole.

1.2.1 Digitization and Metadata

Digitization is the process of converting information into a digital (i.e., computer-readable) form. What this means is very subjective and can depend not just on the underlying type of artifact but also on the purpose of the particular digitization effort. For example, one digitization effort might consist of creating a database with an object identifier and a photo. At the same time, another might copy to a database a catalog index with an object identifier and a short text description. Digitizing an audio recording is very different from digitizing a fabric garment. Nevertheless, in most cases, the effort includes the following steps:

1. Creation of a digital reproduction (e.g., one or several photos, CAT Scan, etc.).
2. Attachment of descriptive, structural, and technical metadata.

Figure 1.6 shows an example of a digitized silk fabric artifact from a digital archive. Digitization efforts face two principal challenges: it is both a time-consuming and costly process. It may require special handling of artifacts and specialized equipment. It certainly requires expert staff.

From the point of view of the CH field, our work fits into digitization efforts by reducing both time and cost of attaching the metadata to the digitized object. Further, our work is meant to facilitate standardization and integration across institutions by making it possible to automatically assign metadata based on the same taxonomy. This is an issue because the field is severely affected by an irregular, limited adoption of cataloging standards. Terminology tends to lack normalization, a problem replicated and aggravated within each national or local language, as well as by the diversity of technical or historiographical approaches to the subjects. Common examples of metadata include the date when the object was made, where it was made, and the materials or techniques used (some of these are shown in Figure 1.6).

Figure 1.6: Example of a digitized artifact from CER.ES, showing a photo, its metadata, and the text description.

Text descriptions of artifacts can be significantly different across archives, the most noticeable difference being a change of language. But this is far from the only noticeable difference. Text descriptions can be different in several other key points:

- Syntax: changes in the arrangement of words, sentence structures, and grammar.
- Length: texts can be short (single sentence) or long (multiple paragraphs).
- Content: the information conveyed in the text, some descriptions might be just about the history of the artifact, others just about its physical properties, and others just about what scene or characters are depicted in the artifact.
- Semantics: Changes in the meaning of words or phrases can significantly impact the interpretation of the text.
- Style: Different writing styles, tones, or voices.

Table 1.1 shows several example text descriptions from different archives. The description from the Victoria and Albert Museum was abbreviated as it would fill more than one page by itself and contains a description of the artistic style, the physical description of the object, the history of the object, the historical context, a summary, and bibliographic references. By contrast, the text description of the artifact from the Israel Museum mentions only an inscription. The dramatic change in the text of the descriptions, combined with changes in language, and changes in the artifacts described, effectively make each archive its own domain from the point of view of NLP.

1.2.2 Literature Analysis and Emotion Detection

In literature analysis, unlike in the analysis of text descriptions of digitized artifacts, the text can be thought of as the artifact itself. Literature analysis of CH refers to the systematic examination, interpretation, and study of literary works, documents, and textual artifacts that hold cultural significance and historical value. It involves the exploration of written materials, such as manuscripts, books, letters, poems, reports, and other forms of written expression, to gain insights into the culture, society, and beliefs of a particular historical period, community, or group of people. Literature analysis plays a crucial role in understanding the past, preserving cultural identity, and uncovering valuable information about the intellectual, social, and artistic aspects of a society. A very good case can be made for defining humans as storytelling animals (Gottschall, 2012). Stories hold a special

Table 1.1: Example text descriptions of digitized silk artifacts in different English language archives.

This silk panel combines the most expensive of materials with the most complex of weaving techniques. (...) These designs are in the idiom we now call Rococo, (...) after the death of Louis XIV (d. 1715) in France (...) and makes bold use of natural history, in particular shells and flowers. (...) Victoria and Albert Museum
Fragment of silk fabric with design of rows of repeating horseshoe motif with flower (peony) base enclosing a flying bird, alternating with stylized flower with crescent leaf in gilt paper strip discontinuous supplementary patterning wefts on a dark green twill-weave silk ground; possibly made for Western market. MFA
Stole-shaped head veil in silk organdie. Oblong shape, made of transparent fabric, the entire surface covered with floral designs. (Homespun silk thread, purchased cotton thread). Length 3410mm, width 500mm. Horniman Museum and Gardens
Hebrew inscription: "Torah Crown" (abbreviated) Israel Museum

place in the hearts of people. Cultures utilize stories to impart social values to successive generations, and the stories that endure from generation to generation are the ones that align seamlessly with the human mind by evoking powerful emotions (GRAHAM & HAIDT, 2012). Feelings and emotions play a fundamental role in the functioning of literature and are central to all narrative (Johansen, 2010; Oatley, 2003). We can think of emotions as a form of metadata humans can infer from a text. Making this "emotional metadata" explicit and readable by computers allows its use in search engines as well as other digital platforms, and facilitates large-scale systematic analysis of literature, such as a study on the intersection between a specific sensory experience and emotion (Massri et al., 2022). Figure 1.7 shows the Smell Tracker [3] dashboard for analysis of emotion in fairy tales as an example of an application that integrates emotion detection and literature analysis.

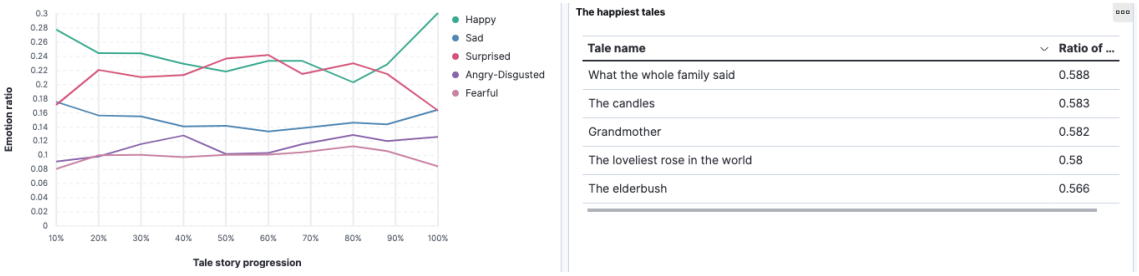


Figure 1.7: Screenshot of the Smell Tracker emotion dashboard.

1.2.3 Digital Platforms

As regards the information about CH, many descriptions and images of objects exist within in-house databases that are only available as local files. In other cases, those records are uploaded by many museums across the globe, in siloed repositories and heterogeneous, often incompatible formats. A few of them give public access to the images and metadata of their objects through APIs, and many more through their websites, but harmonization and integration efforts have been very scarce. Therefore, it is very hard for general audiences, historical experts, and industry to access this knowledge. Some national efforts have aimed to create online portals that aggregate information from CH repositories across the

[3] https://odeuropa.ijs.si/

country [4], but the biggest project in the space is Europeana [5]. It includes digital cultural heritage from over 4000 different European institutions, including a staggering 31M images and 24M texts.

Going beyond the ideas of aggregation and simple search are a series of new digital platforms in the space of Exploratory Search Engines (Palagi, 2018) such as ADASilk (Ehrhart et al., 2021), which can be seen in Figure 1.8, and Smell Explorer (van Erp et al., 2023) . A similar concept is Observatories which integrate data from various domains into a unified exploration, visualization, and analysis interface. One example is Smell Tracker (Massri et al., 2022) which integrates CH olfactory data with olfactory data from other domains, such as scientific publications.

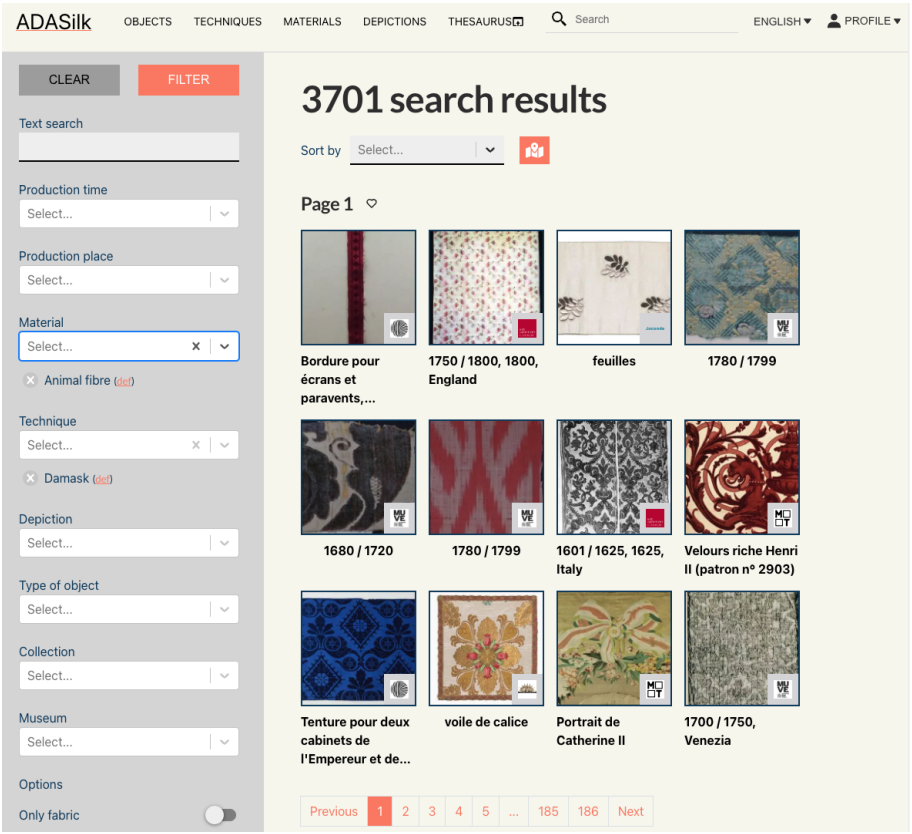


Figure 1.8: Screenshot of the ADASilk exploratory search engine.

In the title of this work, we use the term "Cultural Heritage Archives" to mean digital archives of any kind. That is, any system that stores information about digitized cultural artifacts. Whether this is the website of a GLAM institution, a national or international aggregator, a Knowledge Graph powering an exploratory search engine, or an Observatory. In the case of literature, the archive can be a museum, a library, or a specialized archive like Project Gutenberg [6], a volunteer effort to digitize and archive cultural works.

[4] For example: https://ceres.mcu.es and https://www.pop.culture.gouv.fr
[5] https://www.europeana.eu/en
[6] https://gutenberg.org

1.3 Scientific Context

1.3.1 Broader Context

The work described in this thesis was done in the context of large European research projects.

- "SILKNOW. Silk heritage in the Knowledge Society: from punched cards to big data, deep learning and visual / tangible simulations" (2018) [7] a research project aiming to improve the understanding, conservation, and dissemination of European silk heritage, through advanced computing solutions, matching the needs of diverse users (museums, education, tourism, creative industries, media, etc.), and helping to preserve the tangible and intangible heritage associated with silk. SILKNOW was the winner of the European Heritage Award / Europa Nostra award in the Research Category for 2022. It was funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 769504.
- "ODEUROPA: Negotiating Olfactory and Sensory Experiences in Cultural Heritage Practice and Research" (2021) [8] a research project, bundling expertise in sensory mining and olfactory heritage, aiming to develop novel methods to collect information about smell from (digital) text and image collections with the ultimate goal of showing that critically engaging our sense of smell and our scent heritage is an important and viable means for connecting and promoting Europe's tangible and intangible cultural heritage. This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101004469.
- "MIDAS: Meaningful Integration of Data, Analytics and Services" (2016) [9] a research project aiming to address the needs of policymakers and citizens across Europe by delivering a unified healthcare big data platform to enable better evidence-based long-term policymaking decisions across Europe at regional, national and European levels. This project was supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 727721.

Our work on digitized silk artifacts was enabled by the creation of the SILKNOW Thesaurus (Alba et al., 2022; Léon et al., 2019) which began the process of mapping the metadata of individual archives into a common vocabulary, which would ultimately allow us to create the target labels for the metadata fields we would aim to predict in Rei et al. (2023) and Rei and Mladenić (2020). The source data and mappings were retrieved simply by querying the SILKNOW Knowledge Graph (KG) (Schleider et al., 2021). Our resulting metadata predictions were fed back into the KG which in turn allowed users of ADASilk (Ehrhart et al., 2021) to search and browse digitized artifacts using the predicted metadata, as well as see it directly on the page of the digitized object with an explanation that it was a prediction. Our original work on predicting missing metadata for digitized artifacts from text descriptions in Rei and Mladenić (2020) was originally done in parallel with the work of our colleagues in CV to accomplish the same task from photos (Dorozynski et al., 2019) before coming together in our joint multimodal work: Rei et al. (2023). Our work on literature analysis can be considered as an expansion or continuation of the emotion work done in Massri et al. (2022) with the objective of integrating the predicted emotions into the Smell Explorer (van Erp et al., 2023) similar to the integration of the predicted metadata into ADASilk. Finally, our work in Healthcare Pita Costa et al., 2021 was meant to integrate into a health policy decision support platform (Rankin et al., 2017).

[7] https://silknow.eu
[8] https://odeuropa.eu
[9] http://www.midasproject.eu

1.3.2 Motivation

Cultural heritage fields often harbor knowledge that, while understood by domain experts, remains opaque to the broader public. Despite extensive digitization efforts, these experts frequently encounter difficulty locating specific items in online catalogs, leading to continued reliance on laborious manual consultations. For industry and the broader public access to a lot of CH information is even more challenging. Notably, digitization within the cultural heritage space is far from complete. Many institutions have yet to fully digitize their collections, and existing digitizations often lack crucial metadata, limiting searchability for specific artifacts or the exploration of knowledge associated with CH. Occasionally, "digitization" may simply refer to the creation of a database record for an object, with scant other information available. Records may include only a photo, a brief text description, or partial metadata, without visual or descriptive details. The impetus behind our work is threefold: to aid future digitization through automatic metadata assignment, enhance existing digital collections with supplemental information, and help develop innovative digital platforms that bolster the search, filtering, and browsing experiences for both experts and the public. Emotion detection from text in literature presents interesting challenges, mainly due to the intricate nature of the texts and the nuanced understanding required for emotion labeling, often necessitating expert annotators. Traditionally this has resulted in smaller, imbalanced datasets that contrast with the large, balanced datasets favored by current machine learning practices. Existing works often simplify the rich emotional landscape by using coarse-grained emotion categories and assigning a single label per sentence, potentially overlooking the complex co-occurrence of emotions. The drive towards finer, more comprehensive emotion detection in literature, necessitates innovation to more accurately capture the depth and diversity of human emotions in literary texts. Beyond the applications, we are motivated to explore avenues for addressing some challenges involved in what can be considered the last great frontier for ML: knowledge-intensive resource-scarce problems, where simply getting more labeled data is challenging.

1.3.3 Related Work

This introduction has already provided the reader with the scientific context of our work, and each succeeding chapter will offer its own related work. Here, however, we aim to summarize the most closely related work to our own, primarily from a methodological perspective.

The application of Natural Language Processing (NLP) and text analytics, particularly text classification, within the cultural heritage domain remains relatively limited. The study most closely related to ours is presented in Ruotsalo et al., 2009a, which assessed an Information Extraction approach on 250 manually annotated text descriptions from the Rijksmuseum in Amsterdam. This study successfully extracted key attributes such as Technique, Material, Date, Place, and others—including the Creator of the artwork, style, and subject depicted—from textual descriptions of paintings, achieving an average F1 score of 61.2%. Our approach diverges significantly in several critical respects. Firstly, our classification methodology is more generalizable as it does not strictly rely on the presence of information within the text, thus offering greater resilience to misspellings and non-standard grammatical structures. This is particularly important when attempting to link information known to represent a certain label, such as material types that may be misspelled or specified in uncommon terms in catalog tables. Secondly, our study extends to multilingual datasets from diverse sources, introducing complex challenges that we address through transfer learning. Additionally, our dataset is considerably larger and utilizes automated labeling based on catalog information rather than external annotations.

Regarding multimodal approaches, Belhi et al., 2018 introduced a joint image-text neural network architecture that classifies paintings by artist and year, using a constrained set of textual labels (style, media, and genre) instead of full text descriptions. Another notable work, CLIP-Art (Conde & Turgutlu, 2021), pretrains an image classification model to correlate a concise text vocabulary with images via joint contrastive pre-training, applied to the iMet Collection dataset (C. Zhang et al., 2019) which encompasses labels for visually depicted objects, dimensions, medium, and country of origin. Contrary to these methodologies, our research proposes treating images, multilingual text, and tabular data as equally significant modalities and incorporates data from multiple collections.

In the realm of emotion detection from textual literature, the Tales dataset (Alm et al., 2005) has been a pioneering resource and, until our contribution, it was the only dataset specifically designed for this purpose. It is a small (1,207 examples), imbalanced, single-label multiclass dataset that categorizes texts into just five broad emotional groups derived from Ekman's theory: anger-sadness, fear, disgust, happiness, and surprise. In contrast, our newly introduced dataset is substantially more diverse, comprising 38 labels plus neutral examples with more than 200,000 multilabel samples. Initially, the Tales dataset was employed in Massri et al., 2022 to train an emotion detection model, which was later superseded by our model in the Smell Explorer (van Erp et al., 2023).

1.3.4 Challenges

Cultural heritage as a machine learning application domain is a low-resource, resource-constrained, knowledge-intensive domain. A "Low-resource" domain is an application domain where labeled data is so scarce that it is considered the limiting factor in the performance of ML models. The broader term "Resource-constrained" domain typically accounts not just for data scarcity but also for limitations in computational power, money, or even capable human annotators. One reason for a domain to lack capable annotators is that it is a "Knowledge-intensive" domain. These are domains that require deep domain expertise or specialized knowledge to understand and interpret the data. Common examples include medical diagnosis, legal analysis, scientific research, and engineering fields. For any given task to be performed in CH using ML, there are likely no supervised datasets already built or at most a single small dataset. When creating a dataset for any particular task in CH, even using automatic means to get data from multiple online databases can be challenging due to a mismatch between taxonomies, terminology, cataloging standards, etc., as discussed in Section 1.2.1. Even if this challenge is overcome, the specialization inherent to the task can make the total amount of data low. For example, much of the metadata that is used to describe textile artifacts is necessarily different from that used to describe sculptures. Further, CH data is necessarily limited for many classes: we cannot go back in time, create more artifacts of a certain category and ensure their preservation to the present day. Nor is it likely realistic to significantly increase the number of artifacts in a museum by purchasing or discovering new ones in the context of an ML project. CH is considered a knowledge-intensive domain because it relies heavily on expertise, specialized knowledge, and interdisciplinary approaches. It involves the deep understanding and interpretation of historical and cultural contexts, artifacts, and documents. Experts in cultural heritage often use domain-specific taxonomies and definitions. Accurately labeling artifacts according to these taxonomies is beyond the ability that is reasonable to expect from someone, such as a crowd worker, unfamiliar with them or the artifacts they are meant to categorize. Another example of a knowledge-intensive domain we will use in this work is healthcare. It involves complex medical concepts, procedures, and technologies that require understanding and expertise. Healthcare professionals, such as doctors, nurses, and researchers, undergo extensive education and training to acquire the necessary knowledge

and skills to provide medical care and make informed decisions. Much like CH, healthcare also has its own taxonomies and definitions, and it is also unreasonable to expect laymen to accurately label data in healthcare. Also, like in CH, it might not be feasible to gather additional data in certain cases, such as new or rare diagnoses. Our proposed solutions can be broadly divided into three different, non-mutually exclusive categories: transfer learning, multimodality, and semi-supervised. But even when adopting these proposed solutions, we still face several challenges:

- Transfer learning: use of as much related knowledge as possible to accomplish these tasks.
  - A significant amount of objects in an archive can lack text descriptions.
  - Archives have varying standards for text descriptions of objects, creating significant disparities in content, syntax, and length.
  - Features extracted from intra-archive regularities in objects and text descriptions are not transferable.
  - Geographical spread of archives requires cross-linguality to transfer knowledge from data in different archives.
  - Texts often contain domain-specific terms in specialized areas like the cultural heritage of silk production.
- Multimodality: take advantage of all possible data modalities.
  - Reduced and variable overlap exists between images and text descriptions, with many photographed objects lacking corresponding textual descriptions.
  - Different images may exist for the same object, while multiple text descriptions per object are non-existent.
  - Even drawing from multiple archives, labeled data is relatively small, making it challenging for many multimodal approaches.
- Semi-Supervised: exploit approaches that make use of unlabeled data.
  - Both literature and emotions can be difficult to understand for humans and algorithms.
  - Label noise can be introduced by many semi-supervised processes, such as weak labeling or label propagation.
  - Difficult to identify good negative examples without annotation.

1.4 Hypotheses and Aims

We present an overview of our work as a set of hypotheses defined with specific aims in mind and in relation to prior work. These lead to our contributions in Section 1.5.

Hypothesis 1. Transfer Learning is an effective approach to inferring metadata from text descriptions of digitized cultural heritage objects: Given pairs of text descriptions and metadata (labels) from one archive, we can learn a model that infers metadata from text descriptions in other archives.

- **Aim:** We first need to show that, given labeled data in one archive, we can accurately infer those labels for unlabeled data in that same archive. The next step is to show that we can infer those labels in other archives. Finally, we need to show that this can be done even if the source and target archives are in different languages.

- **Prior state of the art:** The only previous work in predicting metadata in cultural heritage from text descriptions was Ruotsalo et al. (2009b). This approach used extraction, which meant that the labels had to be present in the text descriptions, and the work was only done in a single archive, i.e., no transfer was explored. Their dataset had a total 250 text descriptions and was manually annotated.

Hypothesis 2. Multimodal learning is an effective approach to metadata inference in the cultural heritage domain: Using multiple modalities to infer missing metadata is better than using any one single modality.

- **Aim:** Show that using a multimodal model, incorporating text, image, and tabular classification, is better at inferring missing metadata than any one single modality. We also aimed to introduce a novel dataset to the cultural heritage and multimodal analysis domains.
- **Prior state of the art:** Belhi et al. (2018) present joint image-text neural network architecture for classifying images of paintings by artist and year but their text input consisted of a limited set of labels (style, media, and genre) rather than text descriptions. Clip-Art (Conde & Turgutlu, 2021) had a similar approach, applying CLIP (Radford et al., 2021) to the retrieval of artwork images. It learned to associate a small text vocabulary, akin to labels, with images through joint contrastive pretraining. This was applied to The iMet Collection dataset (C. Zhang et al., 2019), possibly the most similar dataset to our own, it includes images of artworks associated with labels. Another very relevant dataset in this context is Artpedia (Stefanini et al., 2019), a dataset of images of paintings and text descriptions.

Hypothesis 3. Semi-Supervised methods are effective in addressing the challenges of the cultural heritage domain: Using weak labeling, ranking, and propagation we can create an effective dataset for literature analysis without human annotations.

- **Aim:** Generate a large multi-label emotion detection dataset from literature sentences with 38 fine-grained emotions without human annotation and show that it has comparable quality to alternative dataset creation methods by evaluating it using a gold set as well as in transfer learning setting with existing benchmark datasets.
- **Prior state of the art:** The previous emotion detection dataset built from literature sentences, Alm et al. (2005), is single-label with only 5 coarse-grained basic emotions and contains only a total of 1207 sentences. GoEmotions (Demszky et al., 2020) is a social media multi-label emotion detection dataset with 27 emotion labels annotated using crowd workers.

1.5 Scientific Contributions

The core scientific aim of this work is to contribute a methodology that enables ML to tackle resource-scarce knowledge-intensive domains. We focus our work on cultural heritage due to the many challenges posed by this application domain. We also tackle a similar issue with healthcare, although many of the smaller challenges posed by CH are not present. Figure 1.9 shows a flowchart with our methodology, annotated with scientific contributions and our published work. Scientific Contribution 1 corresponds to Hypothesis 1, and so forth.

SC 1. Transfer learning for metadata inference in the cultural heritage and healthcare domains: In Rei and Mladenić (2020), we investigate the ability to use machine learning to infer metadata from text descriptions of silk fabrics. We then show that

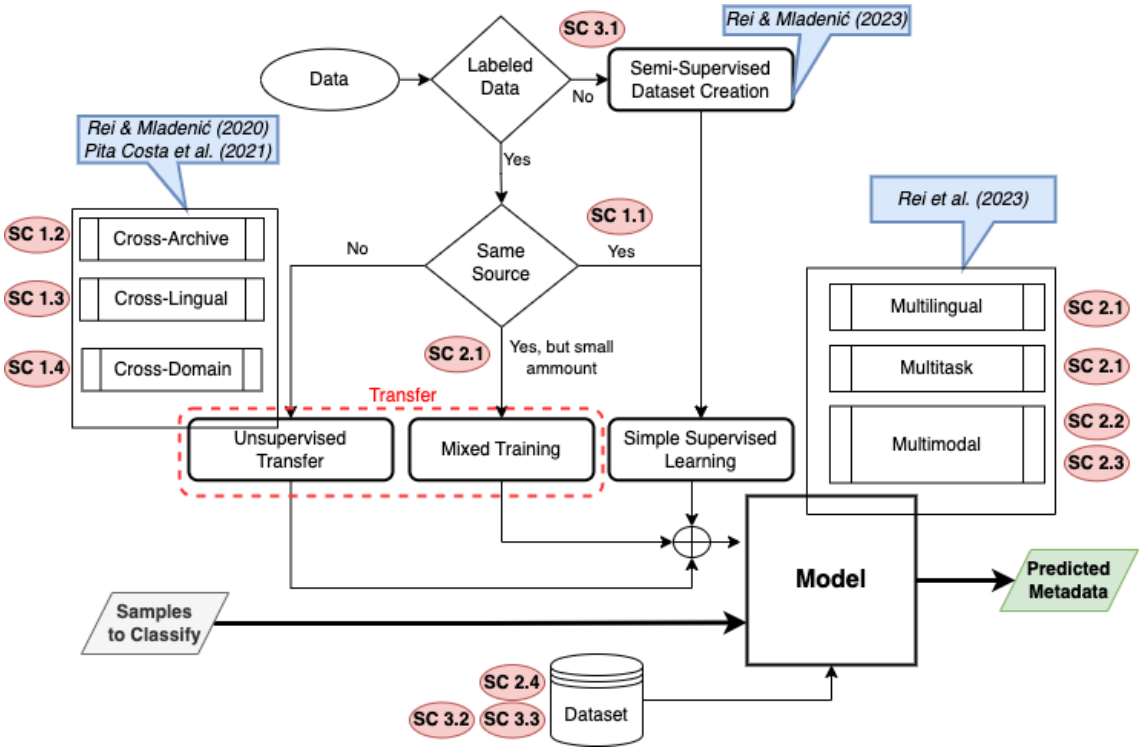


Figure 1.9: Methodology flowchart annotated with specific scientific contributions.

transfer learning works across archives, even when these use different languages. Additionally, in Pita Costa et al. (2021) we also use metadata present in scientific articles to annotate news articles within the healthcare domain automatically. Specifically, we show that:

SC 1.1 Metadata inference from text descriptions of digitized cultural heritage: Important properties that are used as metadata for digitized artifacts can be inferred from text descriptions of those artifacts by using a machine learning approach.

SC 1.2 Cross-Archive transfer: An algorithm can infer metadata for digitized artifacts in one collection, having learned from examples in a different archive.

SC 1.3 Cross-Lingual transfer: An algorithm can infer metadata for digitized artifacts in one collection, having learned from examples in a different archive, even if their languages are different.

SC 1.4 Cross-Domain transfer between medicine articles and news articles: It is possible to leverage metadata in scientific articles to annotate news articles with healthcare topics automatically.

SC 2. Multimodal machine learning for inferring metadata of digitized cultural heritage objects: This group of scientific contributions is derived from Rei et al. (2023). Following our previous work where we showed that it was possible to perform automatic metadata assignment for digitized artifacts from text descriptions (Rei & Mladenić, 2020) and following the work of our colleagues which showed that it was possible to predict the same properties from photographs of the artifacts (Dorozynski et al., 2019):

SC 2.1 Multilingual multi-task metadata inference from text descriptions: We show that we can accurately infer multiple metadata properties for digitized artifacts from text descriptions in multiple archives and languages with a single shared-encoder multilingual multi-task model using mixed training.

SC 2.2 Missing metadata inference from known metadata: We show that using existing metadata in a tabular classification setting can predict missing metadata and provide predictions for records without text descriptions or images.

SC 2.3 Multimodal metadata inference: We introduce a multimodal machine learning approach for predicting the properties of digitized artifacts. Crucially, our approach deals with low overlap between modalities. We show that multi-modality improves the performance of automatic metadata assignment in cultural heritage when compared to single-modality machine learning.

SC 2.4 Multimodal dataset: We introduce a novel dataset to the cultural heritage and multimodal analysis domains that includes data for four different tasks and three different modalities. It consists of harmonized labeled text and image data from heterogeneous, multilingual sources.

SC 3. Semi-Supervised learning for literature analysis in the cultural heritage domain: The previous contributions, dealing with resource-scarce and knowledge-intensive domains, assumed there would be at least some human-labeled examples either in the same domain or in a minimally related source domain. In Rei and Mladenić (2023), we extended our work to its logical extreme, which handles the scenario where no labeled data is available. In tackling this challenge, our main contributions were:

SC 3.1 Semi-Supervised dataset creation: A semi-supervised approach for creating multi-label classification datasets applied to emotion detection from literature.

SC 3.2 A multi-label fine-grained emotion detection dataset for literature: A large, balanced dataset with 38 fine-grained emotion labels, which includes more emotion labels than any other dataset and a more comprehensive taxonomy for emotion detection from text.

SC 3.3 Analysis of our emotion detection dataset: Analysis of emotion correlation and sentiment within the dataset, informing future work in emotion detection.

1.6 Thesis Structure

This thesis started with a brief introduction to the main concepts and context discussed in the next chapters. Chapter 2, associated with SC 1 and published in Rei and Mladenić (2020), focuses on the necessary results for automatic metadata assignment in CH archives. Namely, we show that it is possible to use ML to learn to infer important properties of digitized silk fabrics from the text descriptions of the corresponding objects. It also shows that the knowledge contained in one archive can be used to infer the properties of digitized objects in another archive, even when the text descriptions are written in different languages. Additionally, we also use the same approach to label news articles with healthcare topics, corresponding to SC 1.4 and published in Pita Costa et al. (2021). Chapter 3 leverages the results in Chapter 2 to create a Multimodal Multilingual Multitask model for inferring missing metadata in CH archives. It corresponds to SC 2 and was published in Rei et al. (2023). Chapter 4 corresponds to SC 3 and was published in Rei and Mladenić (2023), it extends the problem of having relatively few labeled examples in or only having examples in a different domain to a situation where no labeled examples are available.

Chapter 2

Transfer Learning

This section discusses the application of transfer learning within expert knowledge domains: cultural heritage and healthcare. In both domains, we utilize controlled vocabularies, encoding expert knowledge and expert and data-guided mapping to supply the labels for training supervised classifiers. Figure 2.1 shows a diagram of the approach described in this section.

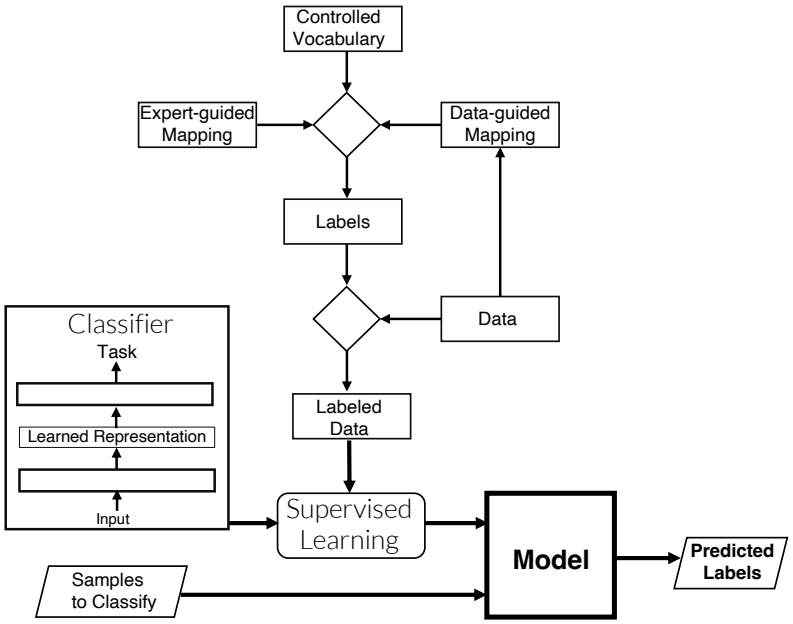


Figure 2.1: Methodological diagram for transfer learning in expert knowledge domains.

In Section 2.1, we employ a thesaurus Alba et al., 2022; Léon et al., 2019, combining its expert-defined hierarchical structure with label counts to standardize labels across various archives. Section 2.2 details our use of Medical Subject Headings, incorporating expert-guided mapping and both pruning of term trees and classifier-guided depth pruning within the hierarchy. Each classifier employs representation learning to facilitate transfer. Section 2.1 describes our use of pre-trained multilingual embeddings, enabling our text classifier to operate across different archives and languages. In Section 2.2, we leverage a pretrained language model to adapt from scientific to news articles.

2.1 Transfer Learning in Cultural Heritage Archive

This section includes the work published in the paper *Cross-Lingual Text Classification for Inferring Variables Describing Silk Fabrics* (Rei & Mladenić, 2020). The paper was presented at the Weaving Europe Silk Heritage and digital technologies conference online in December 2020. This work was then contextualized and incorporated into the article *Open Access to Data about Silk Heritage: A Case Study in Digital Information Sustainability* (Sebastián Lozano et al., 2023), published in the journal Sustainability, Volume 15, issue 19, October 2023. The journal has an impact factor of 3.9 and is indexed in the Science Citation Index Expanded.

We investigated the possibility of inferring metadata from text descriptions of digitized objects and the possibility of transfer learning between different archives and different archives in different languages. The categories of metadata inferred in this work are: "Production Technique", "Material Used", "Production Place" (Country), and "Production Date" (Century). It used an early version of the SILKNOW Thesaurus and some custom mappings to align labels across the different archives and treat them as different domains, evaluating the results of source-only domain transfer. If this baseline works, it is safe to assume that transfer is possible. The dataset used can be seen as an early version of the dataset used in Rei et al. (2023). It consists of text descriptions and metadata of digitized silk fabric objects in the Victoria & Albert Museum (VAM), the Boston Museum of Fine Arts (MFA), and CER.ES (Colecciones en Red). The text descriptions in the first two archives are in English, while the descriptions in the last one are in Spanish. VAM has lengthy, detailed descriptions of every object, while the other two offer predominantly short descriptions. The text classifier used pre-trained cross-lingual aligned word embeddings (Joulin et al., 2018) as its input representation and a Convolutional Neural Network architecture based on the TextCNN (Kim, 2014). The classifier was trained on the VAM text descriptions. When evaluated on holdout data from VAM, its average accuracy was around 94%. When evaluated on the data from MFA, this drops to around 60%, but the drop is not even for all categories of metadata. Predictions for Technique and Material have accuracies of 88% and 78% while Country and Century are much lower at 24% and 48%. The primary cause of this difference is simply that the short MFA text descriptions rarely include relevant information for these properties. The average accuracy for the evaluation on the CER.ES text descriptions was around 55%, with a high of 86% for Country and a low of 21% for Century. The primary reason for the difference was again the presence of relevant data in the text descriptions, with some differences being explained by syntax. From this, we can also infer another specific issue that needs to be explored when considering transfer learning: the choice of source data.

In an earlier experiment, we compared cross-lingual transfer from two different English language source archives, VAM and MFA, to the same Spanish language target, CER.ES. The results are shown in Table 2.1. We observe that transferring from MFA yields better results for Country and Material, whereas transferring from VAM proves superior for Century. This variation in performance can be attributed partially to the similarities in the semantic and syntactic structures of the short text descriptions in both MFA and CER.ES. However, the presence of some place names in CER.ES, which are crucial for accurate classification, and their absence in MFA play a significant role in the results for Country. Additionally, other source factors that likely influence the success of transfer learning include the dataset's size, the frequency of label occurrence, and the extent of label noise. The effects of label noise and dataset size will be further explored in the context of multimodal learning in Section 3.

Overall, this work showed that it was possible to infer properties of digitized silk fabrics from their text descriptions and that transfer learning between different archives was

possible, even when the languages of the archives were different. In terms of application, the method we proposed can be used to fill in missing properties in an archive's categorical description or to align them with a specific vocabulary with the ultimate goal of facilitating the discovery and exploration of silk heritage.

Source	Production Date	Production Place	Material Used
VAM	29.7%	64.2%	30.5%
MFA	20.9%	86.4%	56.4%

Table 2.1: Experimental results on accuracy for cross-lingual transfer from multiple English source archives to the same Spanish target archive, CER.ES. An earlier version of the data was used than in Rei and Mladenić, 2020 hence the difference in results.

Cross-Lingual Text Classification for Inferring Variables Describing Silk Fabrics

Luis Rei[1,2*] and Dunja Mladenić[1,2]

[1]Jožef Stefan Institute, 1000 Ljubljana, Slovenia
[2]Jozef Stefan International Postgraduate School, 1000 Ljubljana, Slovenia

**Corresponding author: Luis Rei (luis.rei@ijs.si)*

ABSTRACT

We present a cross-lingual method for the classification of texts describing silk fabrics with the aim of inferring properties of the fabric. We focus on properties relevant to the heritage context, namely, materials, techniques, production place and date (century). This method can be used to fill missing properties in an archive's categorical description or to align them with a specific ontology with the ultimate goal of facilitating discovery and exploration of silk heritage in general and specifically, across languages. Our text classifier consists of a Convolutional Neural Network (CNN) with pre-trained aligned word embeddings. Our dataset was obtained by crawling web sites of museums in English and Spanish. We classify text descriptions of silk fabrics present in archives into a predefined set of domain-expert defined labels for each property. We evaluate our classifier in different scenarios: trained and tested with data from a single museum; trained with data from one museum and evaluated on a different museum; and finally, trained on data from a museum in one language and evaluated on data from a museum in a different language. Our results vary across scenarios and across properties with above 90% accuracy in certain cases, and around 50% in others.

KEYWORDS: Convolutional Neural Networks, Cross-Lingual text classification, cultural heritage, silk fabrics.

1. INTRODUCTION

A large number of culturally significant historical artefacts have been digitized and made available online. This means that experts in cultural heritage, and often the general public, now have the ability to search for and access information about specific artefacts instantly even when these are stored in distant parts of the world. However, certain challenges to exploring and accessing the digitized information remain, two of which we will address in this work. The first challenge is, obviously, language. For example, the European Union, which includes most of Europe, a continent with a very closely linked cultural history, has 24 official languages. The second challenge is the lack of standardized representation of knowledge across the different archives or catalogues. That is, the lack of a common ontology. Important data, such as the technique used to create a specific silk fabric, is either not specified categorically, or when specified, it does not necessarily use the same term as in a different archive. This difference in terminology has a negative impact on the ability of an expert to find and understand related artefacts across archives. In fact, most online (and offline) catalogues, for each digitized artifact, often have only a title, a short text description of the artefact, an image, and far less often, an incomplete set of arbitrarily defined categorical descriptions in a non-standard terminology. Although not all of these are necessarily present for different artifacts even within the same catalogue. In this work, we use the available text, in whatever language it is written in, both title and short description, to infer categorical properties of the underlying silk fabric that are useful for cultural heritage experts. These properties, which we can variables, can be aligned with a specific ontology or thesaurus which itself can be multi-/cross-lingual such as that presented in (Léon, et al., 2019). When images of the silk fabric are present, it is also possible to infer these properties from them as shown in (Dorozynski, et al., 2019).

We propose a method that infers relevant properties of a silk fabric from a short description of it using supervised cross-lingual text classification based on a Convolutional Neural Network (CNN) and pretrained aligned word vectors (also known as word embeddings). The combination of pretrained vectors and our chosen architecture allows us to obtain good accuracy from a training sets containing from a few hundred to a few thousand examples. We obtained the examples used in training and evaluation via web crawling of specific online catalogues containing digitized silk fabrics.

The scientific contributions of this paper can be summarized as follows. To the best of our knowledge, this is the work in which text classification is applied to inferring properties of silk fabrics from short text descriptions of them. And, by extension, also the first work to tackle this problem in a cross-lingual setting. Which we argue is more useful in the context of a shared cultural heritage. We provide a specific classifier architecture for this task and a brief comparison with other similar approaches. We also articulate some of the challenges that our approach faces. Thus, we elaborate both the task, our proposed solution, and present some of the challenges to improvement in relation to our results.

Following this introduction, in Section 2, we present the details of our data and elaborate the specifics of the task. Section 3 details our approach, in particular, our text classifier, its architecture and how it is trained, as well as relevant hyper-parameters for the experiments. Section 4 consists of our experimental setups and its results and a short discussion of them. Finally, section 5 provides a conclusion.

2. DATASET AND TASK DESCRIPTION

2.1.1. Data Description

Our dataset was obtained by crawling online catalogues relevant to silk cultural heritage. These included the following: Victoria & Albert Museum, London (VAM); Boston Museum of Fine Arts (MFA); and Museos Estatales del MEC (CERES) – a catalogue of multiple museums. Using site-specific human defined rules, only pages containing information regarding silk fabric artefacts were retrieved. From these, the title, text description of the artefact, and any categorical fields present in the HTML were extracted (Figure 1). These categorical fields are extracted and normalized, that is, converted to a standard representation in English, defined by domain experts. The categories used in this work are: “Production Technique”, “Material Used”, “Production Place” (Country), and “Production Date” (century). Each of these categories, referred to in our work as a “variable”, has a predefined set of possible values (“classes” or “labels”) in our dataset given the values present in the original catalogue and the domain experts normalization of them (Table 1).

2.1.2. Data Limitations and Issues

It is important to note the limitations of our data. First, we retrieved data from a very limited set of catalogues (online archives), and thus we cannot expect it to be exhaustive within the silk fabric cultural heritage domain. Second, the procedure of web-crawling often implies certain errors such as the accidental crawling of non-silk fabric artifacts when present in the catalogue such as drawings or books describing silk fabrics but not silk fabrics themselves, or artifacts that include silk fabrics but are not entirely or even principally silk fabrics. Third, our automated normalization process for converting categorical labels present in the different catalogues into a standard set of labels valid across the different catalogues is not perfect. Finally, while conducting this work, we did not possess a perfectly formulated ontology for the silk cultural heritage domain. Thus, with the aid of domain experts, we created our set of categorical properties and labels,

specific to the data we had and our purpose: to show that we can infer meaningful properties from short text descriptions of cultural heritage artifacts.

2.1.3. Task Description

The main task is to infer from a short text description of a silk fabric the values of each category (Table 1). We pose this as a supervised learning task. That is, given a set of texts and respective labels, our classifier learns to infer these variables from the text. Thus, when presented with text without labels, it can "guess" the correct labels (with an accuracy better than a random guess). However, this task can be divided into different sub-tasks, and these further sub-divided, to answer specific research questions. The first sub-division we can obviously conceptualize is that each category, being very different in meaning from the others, can be separated. So, in the context of the work described here, we consider inferring from the text the value associated with its category, a single task. We model the problem as set of four distinct single-task classification problems as opposed to a single multi-task classification problem. While the first division is made from a theoretical modelling perspective, our second division is applicational. We divide the data into three different scenarios: 1) Inferring variables not explicit in records given labelled text examples in the same collection; 2) given labelled text examples from a different collection; and 3) given labelled text examples from a different collection in a different language.

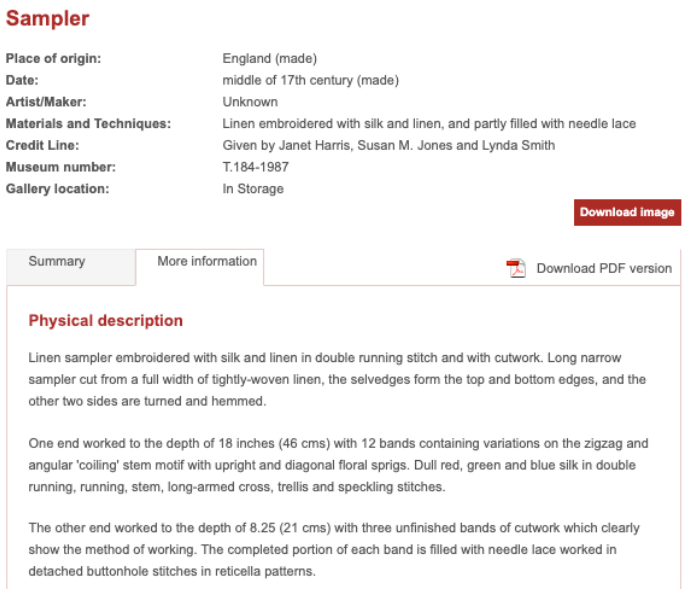


Figure 1. Example of part of a web page from an online catalogue (VAM) containing details on an artefact. Categorical fields can be seen at the top of the image, such as "Place of origin", "Date", and "Materials and techniques".

Table 1. Summary of the data: the variables we attempt to infer, the list of their possible values given our data, and the total number of samples for each variable.

	Technique	Material Used	Production Place	Production Date (century)
Values	Brocading, embroidering, knitting, lace, printing, sewing, velvet, weaving	Cotton, leather, linen, metal_thread, wool, printed, other	Africa, Austria, Azerbeijan, Belgium, UK, China, France, Germany, Greece, Iran, Italy, Japan, Netherlands, Portugal, Russia, Spain, Syria, Turkey, US, South_Asia	10, 14, 15, 16, 17, 18, 19, 20
Number of Samples	3783	4058	8116	7765

3. METHODOLOGY

Our methodology to infer properties of a silk fabric from a short text description of it is based on a supervised text classification approach using a machine learning algorithm. This entails several steps in our implementation: 1) text preprocessing, which converts text into a normalized form and segments it into tokens which mostly correspond to words;

2) the individual words are then converted via a lookup table called an embedding layer into (word) vectors; 3) the vectors are fed into a classifier, which in our case is a Convolutional Neural Network which given an input (the word vectors), outputs a predicted class (1 class out of N possible predefined choices) via a softmax layer. The classifier, being a supervised machine learning algorithm, must "learn" to perform this task via a procedure called "training". In this section, we describe the details of these steps.

3.1.1. Text pre-processing

When using pretrained word embeddings, the main priority of the text-preprocessing step is that it creates word tokens that match those in the word embeddings. Given a text description as a string, the following sub-steps are performed in order:

1- The text string is tokenized using the Moses[1] tokenizer[2] (sacremoses[3] implementation).
2- Case folding i.e., all characters are converted to lower-case.
3- Number replacement i.e., numbers are replaced with whitespace surrounded words, e.g. "1" becomes " one " and 18 becomes " one eight ".
4- String replacements - some common misspellings are fixed and some domain specific words missing from the pre-trained embeddings are replaced with equivalents, e.g., "ssilk" becomes "silk" and "esfumatura" becomes "esfumado".
5- Non-printing characters are removed.
6- HTML elements, resulting from issues with web crawling, are removed, e.g., "kesi" becomes "kesi".
7- Punctuation is removed (note: a single quote used in contractions is preserved).
8- All whitespace characters are converted to the common whitespace (ASCII 32), sequences of multiple spaces are converted to a single space.

3.1.2. Word Embeddings

Continuous word representations, commonly known as word vectors, word embeddings or just embeddings, are a common method of representing text input for natural language processing and machine translation tasks that use Neural Networks as their learning algorithms. These representations can be learned by one algorithm (usually unsupervised) and transferred to a different task or tasks. Word embeddings allow a dense representation of words which include both semantic and syntactical properties. For the purpose of cross-lingual classification, it is advantageous to have a representation where a word in one language is aligned with the same word in a different language. For example, the embedding for the English word "silk" should in some way be similar to the embedding for the corresponding Spanish words "seda". This property allows a classifier to learn from text in a way that is independent of the language of examples shown to it during training. The type of embeddings which possess this property among multiple languages (e.g., English, Spanish, French, …) are called multilingual aligned embeddings. Because of this property, it becomes simple to train a text classifier in one language or set of languages and expect it to work when presented with text in a different language. In this work we use the pre-trained aligned word embeddings described in (Joulin, et al., 2018).

3.1.3. Network Architecture

The architecture of our classifier Neural Network follows from previous work in applying CNNs to text (Collobert, et al., 2011; Kim, 2014). The word embeddings are concatenated and a predefined number of convolutional filters (feature maps) with different fixed window sizes (kernel sizes) are applied to each possible window of words to extract "features". These are then passed through a non-linearity and a max-pooling operation. The idea is to capture the most important feature for each feature map. Pooling over time (1d max pool) deals with variable text lengths - we used a fixed maximum of 300 word-tokens, determined from analysis of the data. After the pooling, the different features for each window are concatenated together, regularized by a dropout layer and put through a final fully connected output layer with a softmax activation to give a distribution of probabilities over the classes (Figure 2). The general intuition behind the algorithm is that each window of size h = 2, 3, 4 learns to extract something similar to word n-gram features where n=h. In this work, the sequence of operations consisting of Convolution, Activation, and Max Pool form a convolutional block. A single convolutional block handles a single window size. We use 3 blocks in parallel, corresponding to the window sizes h = 2, 3, 4. We use the Gaussian Error Linear Unit (GELU) (Hendrycks, et al., 2016) as our activation function and the Alpha Dropout (Klambauer, et al., 2017) variant of dropout. The activation function, dropout variant, dropout probability, convolutional kernel sizes (window sizes), and the number of filters were all treated as hyper-parameters and selected through

[1] http://www.statmt.org/moses/
[2] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl
[3] https://github.com/alvations/sacremoses

hyper-parameter tuning on a subset of the data. In all experiments, the network was trained for 300 epochs, using mini-batch stochastic gradient descent, with a batch size of 64, and an initial learning rate of 0.005.

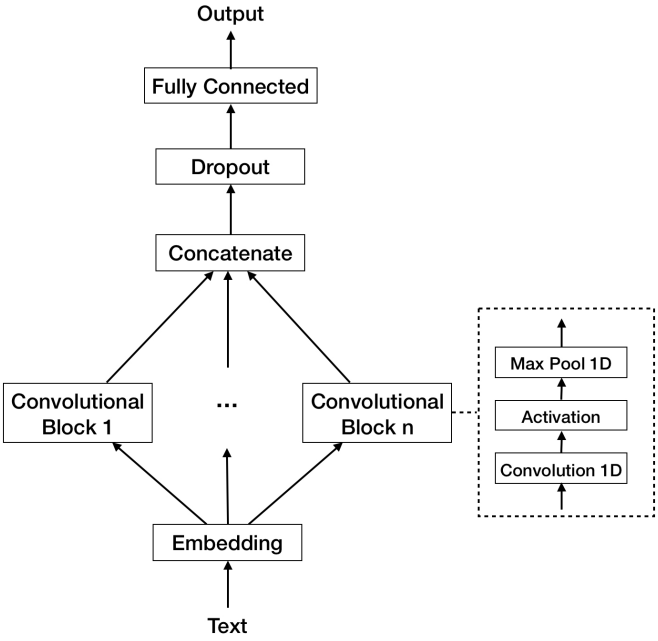


Figure 2. Convolutional Neural Network Architecture for text classification.

4. EXPERIMENTS

4.1.1. Experimental Setup

Our experiments focus on answering specific research questions. Our research questions are posed in terms of specific application scenarios. We begin by presenting the research questions and their application which were originally mentioned in the context of the task description (Section 2.1.3):

1- Given labelled data (digitized artifacts) from one catalogue (e.g., a museum), can we infer those labels (properties) in non-labelled data in **the same catalogue**? The practical applications of this include the ability to infer properties in a catalogue from a subset of that catalogue's data which was semi-automatically or manually labelled, filling missing data, and semi-automatic conversion to a different ontology.

2- Given labelled data from one catalogue, can we infer the labels of non-labelled data **in a different catalogue**? Practical applications include aligning the ontologies of two or more different catalogues, and, if one can be labelled with a standard ontology than, that effort can be leveraged to provide those categorical labels others.

3- Given labelled data from one catalogue, can we infer the labels of non-labelled data **in a different language catalogue**? Applications are the same as in the previous case, but cross-lingual.

For each of these questions, we created a corresponding experimental evaluation scenario. For the first scenario, we use the data we collected from VAM and split it into a separate train and test sets (Scenario 1A). This artificial split of the data was performed using random stratified splitting, a technique which randomly selects the examples in each set while preserving the distribution of the labels from the original set. 80% of the examples are used as training and 20% as test examples. In the second scenario, we used the VAM catalogue as the training set and the MFA collection as the test set (Scenario 2). In the third scenario, we again use VAM, which is in English, for training but we use CERES, in Spanish, as the test set (Scenario 3). Note that we use VAM as a training set in all scenarios purposefully to enable a better comparison between results.

4.1.2. Results and Discussion

Our results (Table 2) clearly show that it is possible to infer properties of silk fabrics from a short description of them, even across catalogues and across languages. The best results are obtained when labelled data from the same catalogue is used (scenario 1). The biggest challenge faced by a text classification algorithm when dealing with short descriptions, in the context of digitized artifacts, is how different these texts are across catalogues, in both form (syntax and length), content (semantics), and in the objects they describe (e.g.,

museums are not random collections of objects but rather curated, often thematically and locally). We can clearly see the difference in results with regards to "production place", and "production date" between scenario 1 and 2. MFA descriptions rarely contain any words relevant to these two properties, while VAM often explicitly mentions regions, cities, and even countries with regards to one and dates with regards to the other. Thus, a text classifier trained on VAM descriptions is ill-prepared to handle MFA descriptions. MFA descriptions, though much shorter than VAM descriptions, usually do include techniques and materials, making the results in scenario 2 for these properties much closer to scenario 1. In the cross-lingual scenario 3 we can see a further performance drop attributable, in part, to the difference in language. Descriptions in CERES, while focused primarily on depictions, often explicitly mention locations (e.g., cities) which helps explain the better-than-expected results for "production place". City names are far easy to align in pretrained embeddings than very domain specific techniques and materials since these embeddings are primarily trained on Wikipedia and aligned through the use of dictionaries that are not domain specific. With regards to dates, when explicit, VAM tends to express them using Arabic numerals and ranges (e.g., "1740-1800") while dates in CERES tend to use roman numerals (e.g., "siglo XIX") which is responsible for part of the difference in accuracy.

Table 2. Evaluation results (accuracy) for the different scenarios.

	Technique	Material Used	Production Place	Production Date (century)
Scenario 1	97.6%	91.4%	97.4%	88.6%
Scenario 2	88.3%	77.7%	24.22%	48.2%
Scenario 3	54.9%	59.8%	86.4%	20.7%

5. CONCLUSION

We have shown that it is possible to infer, from a short text description of a silk fabric, properties relevant in the cultural heritage domain. We have also shown that this is possible in a cross-catalogue and cross-lingual setting. Applications of this development include, but are not limited to, machine-aided improvement of categorical digitized data within an archive, changing ontologies of categorical properties in a catalogue to align them with a different ontology or thesaurus, and helping a centralized resource (e.g., an open knowledge catalogue that includes data from multiple museums) homogenize digitized artifacts across its sources. In this work we've provided a methodology for creating and evaluating a text classifier that can handle this challenge. The source code of the classifier is available online under an open source license[4]. Several interesting avenues of future work remain open, especially along domain adaptation. For example, it would be interesting to have pretrained embeddings that are more tuned to the cultural heritage domain. The challenge to this lies in obtaining sufficient amount of text to perform such adaptation. A second avenue would be the adaption of the multilingual alignment to include the use of such data as well as domain-specific dictionaries such as the thesaurus we mentioned previously..

6. ACKNOWLEDGEMENTS

This work was funded by the European Union's Horizon 2020 research and innovation program, "SILKNOW. Silk heritage in the Knowledge Society" project, grant agreement No. 769504.

7. REFERENCES

LÉON, A., GAITÁN, M., SEBASTIÁN, J., PAGÁN, E. A., and INSA, I., 2019. SILKNOW. Designing a thesaurus about historical silk for small and medium-sized textile museums. In *Science and Digital Technology for Cultural Heritage-Interdisciplinary Approach to Diagnosis, Vulnerability, Risk Assessment and Graphic Information Models: Proceedings of the 4th International Congress Science and Technology for the Conservation of Cultural Heritage (TechnoHeritage 2019)*, Pilar Ortiz Calderón, Francisco Pinto Puerto, Philip Verhagen, Andrés J. Prieto (Eds.), p. 187, CRC Press, Sevilla, Spain.

DOROZYNSKI, M., CLERMONT, D., and ROTTENSTEINER, F., 2019. MULTI-TASK DEEP LEARNING WITH INCOMPLETE TRAINING SAMPLES FOR THE IMAGE-BASED PREDICTION OF VARIABLES DESCRIBING SILK FABRICS, In

[4] SilkNOW Text Classification : https://github.com/silknow/text-classification

ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci., IV-2/W6, D. Gonzalez-Aguilera, F. Remondino, I. Toschi, P. Rodriguez-Gonzalvez, and E. Stathopoulou (Eds.), pp. 47–54, Copernicus Publications, Göttingen, Germany.

JOULIN, A., BOJANOWSKI, P., MIKOLOV, T., JÉGOU, H. AND GRAVE, E., 2018. Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Ellen Riloff, David Chiang, Julia Hockenmaier, Jun'ichi Tsujii (Eds.), pp. 2979-2984, Association for Computational Linguistics, Brussels, Belgium.

COLLOBERT, R., WESTON, J., BOTTOU, L., KARLEN, M., KAVUKCUOGLU, K., & KUKSA, P., 2011. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12: 2493-2537.

KIM, Y., 2014 Convolutional Neural Networks for Sentence Classification, In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Alessandro Moschitti, Bo Pang, Walter Daelemans (Eds.), pp. 1746-1751, Association for Computational Linguistics, Doha, Qatar.

HENDRYCKS, D. AND GIMPEL, K., 2016. Gaussian error linear units (gelus). *arXiv preprint* arXiv:1606.08415.

KLAMBAUER, G., UNTERTHINER, T., MAYR, A., & HOCHREITER, S., 2017. Self-normalizing neural networks. In *Advances in neural information processing system*s, Ulrike Von Luxburg, Isabelle M Guyon, Samy Bengio, Hanna M Wallach, Rob Fergus (Eds.), pp. 971-980, Curran Associates Inc., New York, United States.

2.2 Transfer Learning from Science Articles to News Articles

This section includes the work published in the article titled *NewsMeSH: A new classifier designed to annotate health news with MeSH headings* (Pita Costa et al., 2021). The article was published in the Artificial Intelligence in Medicine journal, Volume 114, April 2021. The journal has an impact factor of 7.5 and is indexed in the Science Citation Index Expanded. We presented an automated text classifier based on MeSH thesaurus. The mapping from MeSH, which was designed to catalog scientific papers, to news articles followed 3 different lines: 1) domain expert-guided mapping; 2) pruning of irrelevant subtrees such as geographic location; 3) automatic mapping which analyzed the output of the classifier to determine the best depth of the MeSH tree to use. Thus, although the actual domain transfer is source-only domain transfer, the labels are partially determined in the target domain. We trained two different classifiers, one a nearest centroid classifier (Henderson, 2009; Mladenić, 1998) and the other based on a pre-trained neural network, XLNet (Yang et al., 2019) [1]. The training data consisted of nearly 28M scientific article abstracts and the evaluation data consisted of 325K held-out scientific articles and 100 news articles in 5 separate health domains: Mental Health, Diabetes Mellitus, Coronavirus, Childhood Obesity, Children in Care. The news articles were manually annotated by health professionals with experience in using the MeSH vocabulary. We used the nearest centroid classifier to determine the optimal MeSH tree depth cut-off for categorizing news articles, as switching the depth of the labels used does not require any extra training and thus is very quick to do. With a fixed depth, the problem is a multi-label classification problem. When evaluating in-domain, the NN achieved a superior macro-averaged F1 score of 56% compared to the 44% of the nearest centroid classifier. But when evaluating news articles, the nearest centroid classifier managed a slightly higher 38% F1 compared to 36% for the NN. We have clearly shown that this approach is viable, but the significant drop in F1 scores when comparing in-domain F1 to out-of-domain F1 scores suggests the need to explore unsupervised domain adaption techniques. In terms of application impact, we proposed an innovative approach to advanced search, exploration, and analysis of health-related news to support the workflow of health professionals, particularly those involved in policymaking. Basing the approach around the use of the MeSH vocabulary means that health professionals can use a vocabulary they are already familiar with to explore and derive insights from news events.

[1] The classifier code is available online at: https://github.com/lrei/newsmeshexp

Artificial Intelligence In Medicine 114 (2021) 102053

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

NewsMeSH: A new classifier designed to annotate health news with MeSH headings

Joao Pita Costa [a,b,*], Luis Rei [a], Luka Stopar [a,b], Flavio Fuart [a,b], Marko Grobelnik [a,b], Dunja Mladenić [a,b], Inna Novalija [a], Anthony Staines [c], Jarmo Pääkkönen [d], Jenni Konttila [d], Joseba Bidaurrazaga [e], Oihana Belar [e], Christine Henderson [f], Gorka Epelde [g,h], Mónica Arrúe Gabaráin [g,f], Paul Carlin [i], Jonathan Wallace [j]

[a] Jožef Stefan Institute, Slovenia
[b] Quintelligence, Slovenia
[c] Dublin City University, Ireland
[d] University of Oulu, Finland
[e] BIOEF, Spain
[f] Northern Ireland Department of Health, UK
[g] Vicomtech Foundation, Basque Research and Technology Alliance (BRTA), Spain
[h] Biodonostia, Spain
[i] Open University, UK
[j] Ulster University, UK

ARTICLE INFO

Keywords:
Big data
Semantic technologies
Public health
Healthcare
Text mining
MeSH headings
MEDLINE
PubMed
COVID-19
Diabetes
Mental health

ABSTRACT

Motivation: In the age of big data, the amount of scientific information available online dwarfs the ability of current tools to support researchers in locating and securing access to the necessary materials. Well-structured open data and the smart systems that make the appropriate use of it are invaluable and can help health researchers and professionals to find the appropriate information by, e.g., configuring the monitoring of information or refining a specific query on a disease.
Methods: We present an automated text classifier approach based on the MEDLINE/MeSH thesaurus, trained on the manual annotation of more than 26 million expert-annotated scientific abstracts. The classifier was developed tailor-fit to the public health and health research domain experts, in the light of their specific challenges and needs. We have applied the proposed methodology on three specific health domains: the Coronavirus, Mental Health and Diabetes, considering the pertinence of the first, and the known relations with the other two health topics.
Results: A classifier is trained on the MEDLINE dataset that can automatically annotate text, such as scientific articles, news articles or medical reports with relevant concepts from the MeSH thesaurus.
Conclusions: The proposed text classifier shows promising results in the evaluation of health-related news. The application of the developed classifier enables the exploration of news and extraction of health-related insights, based on the MeSH thesaurus, through a similar workflow as in the usage of PubMed, with which most health researchers are familiar.

1. Introduction

The day-to-day growth of knowledge to support public health and healthcare available online has reached a volume that is very hard to assimilate when researching specific health-related topics. Evidence of this abundance of information is the open scientific biomedical knowledge base – MEDLINE – and its comprehensive controlled vocabulary – the Medical Subject Headings (MeSH) thesaurus – which facilitates the correct refinement of a PubMed search based on article metadata [30]. Aiming to address this need, we have: (i) developed a novel text classifier trained on the existing hand annotations of MeSH heading classes given to biomedical research papers in MEDLINE; (ii)

* Corresponding author at: Jožef Stefan Institute, Slovenia.
E-mail address: joao.pitacosta@ijs.si (J. Pita Costa).

https://doi.org/10.1016/j.artmed.2021.102053
Received 6 April 2020; Received in revised form 21 January 2021; Accepted 11 March 2021
Available online 13 March 2021

proceeded with an extensive evaluation, both on new scientific articles, and on news articles over trending health topics; and (iii) provided a useful and easy-to-use web interface to access the MeSH classifier, as well as an API to allow its integration in existing systems.

The MeSH classifier approach discussed in this paper provides automated annotations of any text, and is thus useful to identify insight in health-related text as, e.g., reports, articles or news. To validate the proposed approach and the constructed classifier, we focus on news annotation. To that end, the integration of the MeSH classifier with a news engine allows the usage of MeSH classes on queries, and for visualisations to explore the health subtopics of interest (see Fig. 1 for illustration). Moreover, this integration enables health professionals to use the MeSH controlled vocabulary with which many are familiar with, to enrich and extend their workflow when exploring and monitoring worldwide news.

1.1. Motivation

With the accelerating use of big data, and with the analysis and visualisation of this information being used to positively affect the daily lives of people worldwide, health professionals need efficient and effective technologies to derive meaning and knowledge from information outputs, when planning and delivering care. The growth of online knowledge requires that the information sources utilised are complete, of high-quality and accessible. A particular example of this is the COVID-19 outbreak [39] that motivated worldwide joint initiatives to help monitor the disease (as [2] and [34]), and understand it better [20], including the crowdsourcing initiative to build new machine learning methods based on biomedical knowledge [13]. In the context of these global initiatives, the proposed text classifier was designed to automatically annotate text. Its development was motivated by the potential for the classification of health reports and news articles on Coronavirus, Mental Health and Diabetes, taking into consideration the pertinence of the further knowledge on the disease and the virus itself, but also the known relations with diabetes [9] and the impact of the social distancing it is generating on mental health [40].

A particular example of a well-established, useful and meaningful tool in the daily life of health professionals is the PubMed search engine, which allows access to state-of-the-art medical research. This tool is frequently used to gain an overview of a certain topic using several filters, tags and advanced search options. PubMed has been freely available since 1997, providing access to references and abstracts on life sciences and biomedical topics. MEDLINE is the underlying open database [16] served by the controlled vocabulary of the MeSH Headings, both of them maintained by the North American National Library of Medicine (NLM). The MeSH vocabulary is often used by health professionals to refine the search results provided by PubMed. This is done via the PubMed search engine directly, or via research assistant tools that integrate the access to this vocabulary such as Zotero.

The gain of automated knowledge discovery from MEDLINE/MeSH is transformative in medical research and can influence the progress of biomedical research [37]. In the context of the meaningful integration and usage of data, the EU H2020 project MIDAS (Meaningful Integration of Data, Analytics and Services) [29] is developing a big data platform that facilitates the utilisation of healthcare data beyond the existing isolated systems, making that data available for enrichment with open data. This data fusion approach thus enables evidence-based health policy decision making, and potentially may lead to significant improvements in healthcare and quality of life for all citizens [4].

The proposed classifier can also be easily integrated into a news search engine. There are several examples of such systems, and a range of news sources that can be annotated by the classifier to leverage its potential. The worldwide health monitoring potential of this tool was discussed in [27] in the context of Public Health decision-making support, though its application can extend to any domain where the automated annotation using terms from a health-related vocabulary such as that of the MeSH thesaurus could be useful.

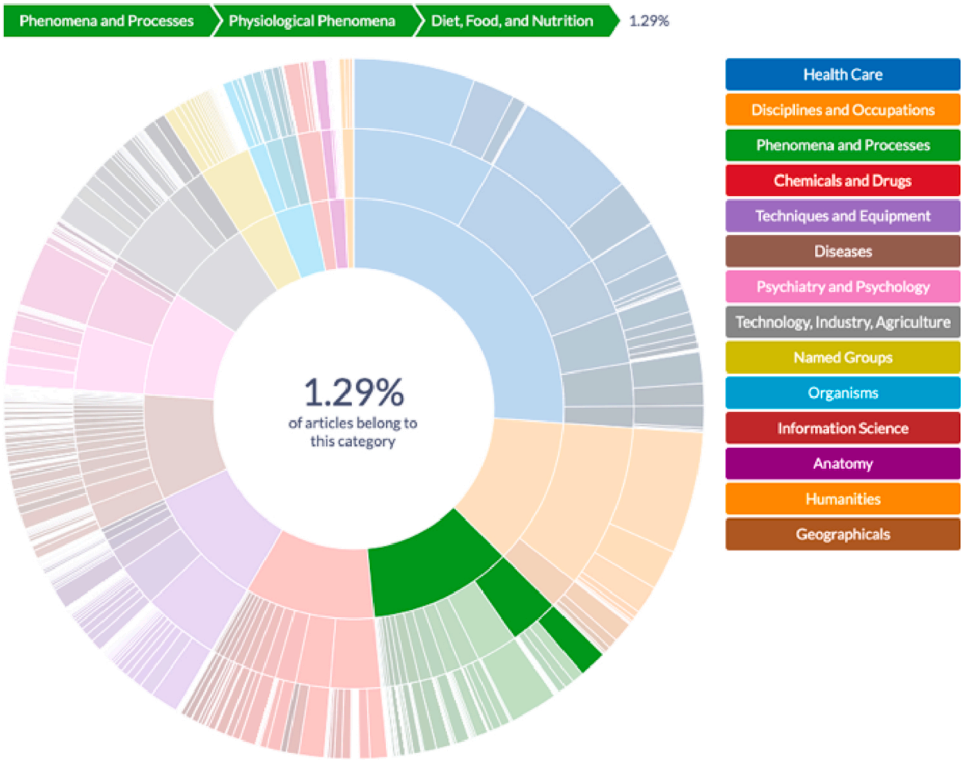


Fig. 1. A potential impact of the MeSH classifier applied to the classification of news: the percentage of news published in 2018/2019 that are annotated with the MeSH class "Public Health".

1.2. Related work

There have been several well-accepted initiatives to use MeSH for the classification of text, often focusing on specific scientific problems [33]. More general approaches include the Medical Text Indexer (MTI) [1], that provides MeSH indexing recommendations to support the human indexers of the NLM using k-nearest neighbors (KNN), pattern matching and indexing rules, reaching a F measure of 56.37 %; and the MeSH Now [19], including a learning-to-rank framework achieving a F measure of 61 %. The latter two are very different in characteristics from the MeSH classifier proposed in this paper: the MTI and the MeSH Now are tools to address the needs of the PubMed user. The NLM also made available other useful tools like, e.g., the MeSH on Demand [32], that suggests MeSH vocabulary explicitly mentioned in the input text; and the Semantic MEDLINE [14], that aims for the semantic knowledge representation of MEDLINE itself. The Semantic MEDLINE initiative is different in the sense that it has focus on research on the semantic knowledge specific problems related to the summarization of MEDLINE citations. The MeSH Now technology is also aiming for the automatic MeSH indexing, with a tailor-fit methodology to the MEDLINE articles. The MeSH classifier we are proposing is also trained over the knowledge base provided by MEDLINE, integrating the concerns from the public health community. Though, it is a hierarchical labelling method, designed to perform efficiently in text that is not tied to the formal jargon of the domain, allowing for the classification of, e.g., reports or news articles. Other alternatives include the MeSHLabeler [17], that combines predictions from MeSH classifiers, KNN and pattern matching, MTI and correlations between MeSH terms to achieve a F measure of 62.48 %; and the DeepMeSH [24] that incorporates deep semantic information and the 'learning to rank' framework of MeSHLabeler to reach a better F measure of 63.23 %. FullMeSH [7] is a recent alternative that is leveraging the increased availability of full text scientific articles in MEDLINE, combining semantic frameworks to get deeper insight from the paper's content in each section, achieving a F measure of 66.76 %, higher than DeepMeSH and MeSHLabeler. Other MeSH-based classifiers to take into account are the end-to-end model AttentionMeSH [12], using deep learning and attention mechanism to index MeSH terms to biomedical text, to achieve a F measure of 68.4 % with a high level of interpretability; the MeSHProbeNet [41], deep learning and self-attentive MeSH probes to index with achieving an F measure higher than DeepMesh and AttentionMeSH, reaching 67.89 %; and the BERT-MeSH [42], a pre-trained deep contextual representation, BERT (Bidirectional Encoder Representations from Transformers), capturing deep semantics of full text, reaching a F measure of 69.2 %. The related work also includes other approaches: the BioNLP (biomedical natural language processing) [3], that is a biomedical text mining combining controlled vocabularies and ontologies available through the National Library of Medicine's Unified Medical Language System (UMLS) [5] and MeSH with the conventional one-vs.-rest (OVR) classification setup; the semantic biomedical question answering system SemBioNLQA for retrieving information based on natural language questions [31]; the usage of topic models to enrich the meta information provided by MeSH [23]; and the foundational work on the medical indexing expert system, MedIndEx, using knowledge-base frames to guide indexers in completing indexing frames [11].

2. Building the classifier

2.1. Data and metadata description

The 2019 version of the MEDLINE dataset used to build the automated classifier proposed in this paper includes citations from more than 5,200 journals worldwide in approximately 40 languages (about 60 languages in older journals). It stores structured information on more than 27 million records dating from 1946 to the present. In most cases, the title and abstract are available but not the main body of the work. About 500 000 new records are added each year. 17.2 millions of these records are listed with their abstracts, and 16.9 million articles have links to full-text, of which 5.9 million articles have full-text available for free online use (see Table 1 for more information). In particular, it includes more than 43 thousand articles with the key-words string "public health".

The comprehensive controlled vocabulary associated with the MEDLINE dataset – MeSH – delivers a functional system of indexing both journal articles and books in life sciences. It has proven very useful in the search of specific topics in medical research and is commonly used by medical researchers conducting initial literature reviews before engaging in particular research tasks [28]. Trained NLM librarians annotate the articles in MEDLINE with MeSH descriptors. These descriptors permit the user to explore a certain biomedical related topic, which relies on curated information made available by the NLM. MeSH is composed of 16 major categories (covering anatomy, diseases, drugs, etc.) that further subdivide from the most general to the most specific, with as many as 13 hierarchical depth levels.

This rich data structure in the MEDLINE open set is manually annotated (although assisted by semi-automated NIH tools) and therefore is not available for the most recent citations. The MEDLINE dataset is mostly in English but also includes a significant volume of abstracts translated from other languages.

2.2. Nearest centroid classifier

We have made available an automated classifier inspired by [21] and based on [10] that is able to suggest MeSH categories for any health-related text. It is trained with the part of the MEDLINE dataset that is already annotated with MeSH and is able to suggest categories for submitted text snippets. These texts can be abstracts that do not yet include MeSH classification, medical summary records or even health related news articles.

To have an efficient automated annotation of text based on the already existing health-related categories provided by MeSH, we use the nearest centroid classifier [22] constructed from abstracts of the MEDLINE dataset and their associated MeSH annotations. Each document is embedded in a vector space as a feature vector bag-of-words (BoW) of Term Frequency-Inverse Document Frequency (TFIDF) weights [18]. The features are words and phrases and the weights reflect how important a word or a phrase is to a document within the collection. The classifier is trained in such a way that for each category, a centroid is computed by averaging the embeddings of all the documents in that category. For higher levels of the MeSH structure, as suggested in [27], we also include all the documents from descendant nodes when computing the centroid. To classify a document, the classifier first computes its embedding and then assigns the document to one or more categories whose centroids are most similar to the document's embedding. We measure the similarity using pairwise cosine distance (BoW Similarity Measure) between the embeddings. Fig. 3 shows the architecture of the MeSH Classifier, where the BoW Vocabulary is calculated on the whole collection of abstracts during the process of training the classifier and is used for embedding documents in a vector space. BoW Similarity Measure is used during the process of classification, when documents are being annotated by MeSH categories.

Table 1
Dataset description based on the statistics for the open dataset MEDLINE and the MeSH headings.

Items	Public Health	All Domains
Number of abstracts	110023	27361292
Number of full-text articles	43844	17538890
Number of languages	42	58
Number of MeSH heading descriptors	10756	29256
Maximum depth of the MeSH tree	6	13
MeSH tree roots (major categories)	3	16

The classifier checks all the categories and returns an ordered list of all the MeSH categories according to their relevance for annotating the provided text. The demonstrator version of the MeSH classifier available online through a web portal[1] (shown in Fig. 2) provides the position number in the annotation and the percentage representing the weight of the MeSH term in the annotation (based on the cosine similarity). The classifier can be also used through a REST API[2], using a POST call, and taking a JSON input that includes the text to be classified. The availability of a well-defined API facilitates further integration with news aggregators (discussed in Section 4) and other systems. The MeSH classifier can suggest categories for any text documents including research papers, medical reports and news articles.

2.3. Large pretrained transformer classifier

Parallel to the Nearest Centroid classifier we implemented a text classifier based on XLNet, a large pretrained transformer model inspired by BERT [8]. We used the XLNet-Base cased variant with 12-layers, hidden size of 768, 12 attention heads and totalling 110 M parameters which matches the model parameters of the base BERT model. We chose XLNet over BERT simply because it has performed better on several text classification tasks. Our model adds a linear layer following the final hidden state corresponding to the [CLS] token (commonly used as the aggregate sequence representation). The model is then fine-tuned on our data using Binary Cross Entropy (BCE) loss. This setup is similar to the original BERT and XLNet papers except for the use of BCE loss, as in this work, the end task is framed as a multilabel task rather than simply multiclass. This is also different from the Nearest Centroid classifier which is trained as a fully hierarchical model. While the Nearest Centroid classifier was used for several experiments and explorations, the Large Pretrained Transformer classifier drew on some of those, such as a fixing label depth at 3, and is presented here only as a comparison to show the potential of the results that can be achieved when using state of the art text classification approaches.

2.4. Learning approach

Research showed there is no existing database of curated and reliable data that can be used as the golden standard for testing automatic MeSH classifiers. For this purpose, we evaluated the system by leaving out one year of MEDLINE abstracts to be used as an evaluation dataset. For this reason, the classifier was trained on the MEDLINE 2018 dataset (27837540 article abstracts), leaving the latest batch of annotated abstracts out. Then, the model was evaluated against new data from the MEDLINE 2019 dataset (325128 articles), i.e., the hand-annotations of the articles that were not annotated in the 2018 version. Both of the MEDLINE dataset versions are accessible at [35], while the MeSH data including the MeSH tree is available at [36]. For the sake of the coherency of the evaluation, the new MeSH descriptors in the 2019 version that are not present in the 2018 version were not considered in the comparison between the automated annotation and the hand annotation. The complexity of the MeSH tree impacts the learning procedure in the time spent to this aim.

To improve the learning time, we reduced the complexity of the MeSH tree by deleting some of the branches that (i) are loosely related to the public health and healthcare topics (e.g., *Geographicals*); (ii) that can be better classified with other taxonomies (e.g., *Information Science*); or the refer to the bibliographical details (e.g., *Publication Characteristics*). Preliminary tests showed that the results of the automated classification of health-related text snippets were not impacted by this reduction of the MeSH classes considered. Though, the learning time was much improved with the mentioned reduction of the MeSH tree complexity.

[1] https://qmidas.quintelligence.com/classify-mesh-major/
[2] https://qmidas.quintelligence.com/classify-mesh-major/api/classify

3. Evaluation

3.1. Evaluation methodology

3.1.1. Evaluation over research papers

The main goal was to determine the performance of the MeSH nearest centroid classifier for medical and news texts. Additionally, the evaluation results should provide an estimate for an optimal similarity cut-off, classification depth and a decision regarding the classification of all MeSH terms or major classes only. In the case of the Large Pretrained Transformer classifier, we don't do any hyper parameter tuning and instead use the parameters chosen in the original BERT paper for its evaluation on the GLUE benchmark tasks [38] with the initial learning rate of 3e-5.

3.1.2. Evaluation over news articles

To guarantee the coherence of the evaluation, we used an adaptation of the evaluation described in Section 3.1.1., but in this case we have the domain experts annotating articles selected in the context of the health domains corresponding to their areas of expertise. Each of the five experts provided between MeSH annotations to news articles on topics related to their expertise. This dataset of hand-annotated news articles, as well as the results in full of the evaluation of the MeSH classifier over these are available at [26]. The dataset was built by collecting news articles that relate to the five topics selected, and are distinguished by these, including the article title and body, as well as the assigned MeSH heading identifiers that can be used to consult more information on each of them through PubMed. This allowed us to evaluate our classifiers in these specific health topics (e.g., diabetes or mental health).

3.1.3. Adopted evaluation approach

In our evaluation approach, we use the following measures: precision, recall, and F1-score. In our case, precision is the fraction of detected MeSH terms that are relevant for a specific article. Recall is the fraction of the relevant MeSH terms that are successfully detected by the classifier, i.e., the number of correct classes divided by the number of classes that should have been returned. To compute these values in case of the centroid classifier, we choose a cutoff similarity and associate the document with all categories that have a higher similarity. The best performing cutoff similarities differ slightly in different domains, but are generally around 0.2. What we consider as "correct" annotations are the terms that were manually assigned by the experts. The precision and recall in terms for Type I and Type II errors can be expressed through the descriptions in the Table 2. A combination of precision and recall is provided in the form of F1-score. F1-score is the harmonic mean of precision and recall. Note that, to proceed with the evaluation of the system we left out one year of MEDLINE abstracts to be used as an evaluation dataset.

3.2. Nearest centroid classifier tuning

The evaluation results provide an estimate of an optimal similarity cut-off, and classification depth. We start from the evaluation of the classifier against annotated scientific articles, and then elaborate on the evaluation of the classifier against annotated news articles.

3.2.1. Determine optimal similarity cut-off

The classifier provides a weight value for each label. We would like to determine what would be the optimal cut-off to be used for reporting the predicted labels. In principle, by decreasing the cut-off value we increase the precision and decrease the recall. In Fig. 5 we show the value of F1 for different cut-off thresholds ranging from 0 to 0.9 with step 0.1.

3.2.2. Determine optimal tree depth for classification

We perform our evaluation by comparing matches at different tree

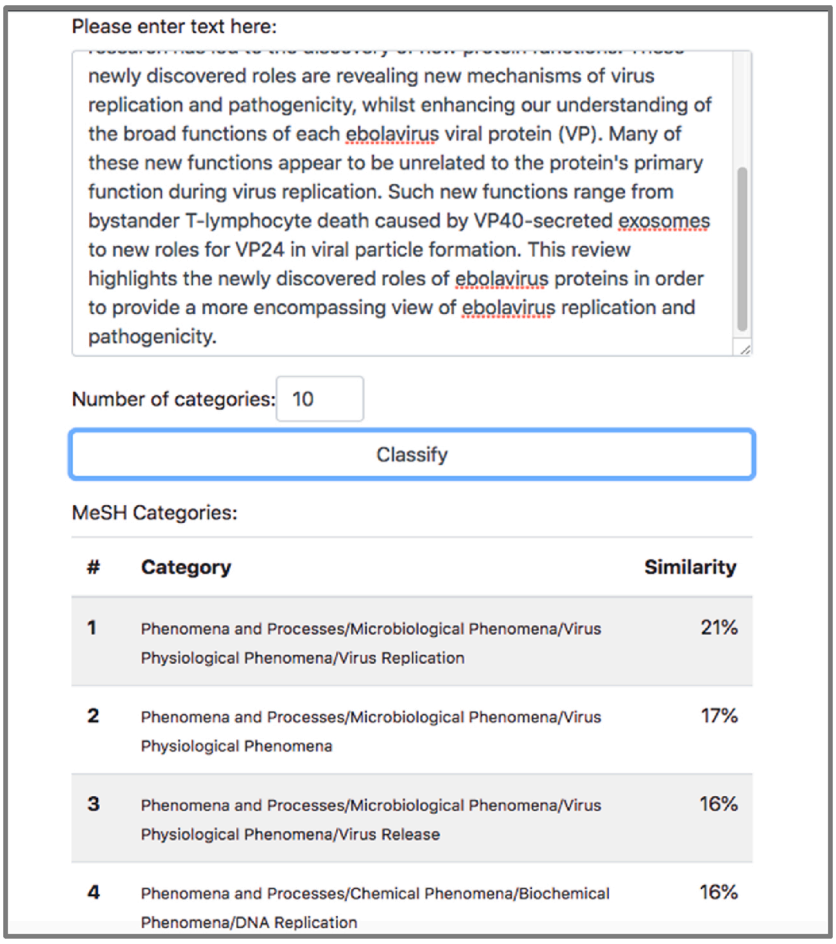


Fig. 2. An example of the MeSH classifier output for the automated MeSH annotation of a scientific article abstract extracted from PubMed (in the body of text above), the MeSH class rank (on the left), their MeSH tree path in the MeSH ontology-like structure (in the centre), and a measure of the similarity of each class to the text (on the right).

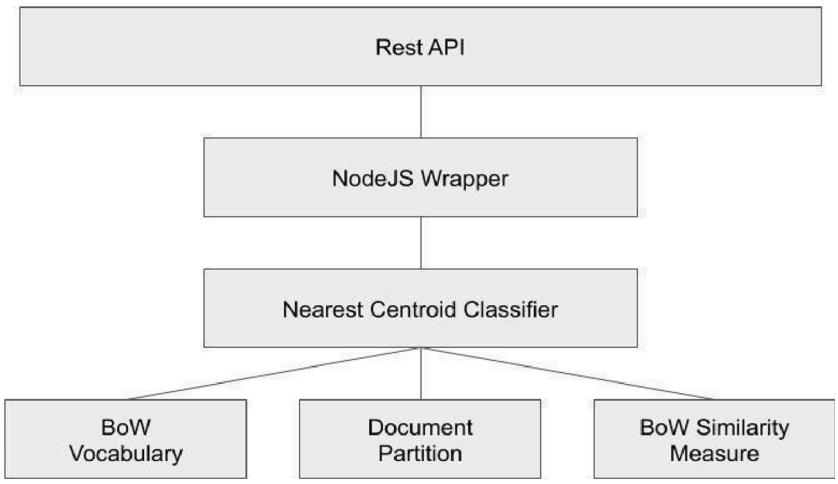


Fig. 3. A high-level diagram of the MeSH Classifier architecture.

depths. The reason is that the aim of the classifier is to assist experts in detecting broader topics, and not find exact MeSH term matches. As the system has been trained on abstracts only, we find that exact matches would be difficult to achieve. Another reason for aiming at broader topics is the intended usage of the classifier on non-medical articles (e.g., news), where detecting those broader topics is a pragmatic goal.

3.3. Evaluation results for setting the parameters on the annotation of MeSH classes over scientific papers

3.3.1. Evaluation setting

We have used the experimental evaluation to determine whether only the major MeSH classes should be returned by the classification model or all the relevant MeSH classes. A major MeSH class is a MeSH

Table 2
Description of the variables used in MeSH classification.

	Meaning and description		
TP	True Positive	correctly identified	Number of detected MeSH classes matching the manual annotation.
FP	False Positive	incorrectly identified	Number of detected MeSH classes not matching the manual annotation.
TN	True Negative	correctly rejected	Number of not detected MeSH classes that are also not in manual annotation. This measure is not needed in the F1 calculation.
FN	False Negative	incorrectly rejected	Number of manual annotations that were not detected by the MeSH classifier.

descriptor, which is viewed as a focus of the paper, while minor MeSH classes are mentioned in the paper. For example, in a paper on survival following myocardial infarction in Ireland, myocardial infarction would be a major term, and Ireland a minor term, as the focus of the paper is on survival, and not on the locations where the patients lived.

3.3.2. Evaluation results

The diagrams of Fig. 4 compare the evaluation results based on the recall (colour code) over the similarity cut-off threshold (Y-axis) and MeSH tree depth (X-axis), for major MeSH classes with those for all MeSH classes. As expected, higher cut-off yields higher precision results and lower recall results. Lower depths yield both better precision and recall. Roughly, for the depth 3 in the MeSH tree, over cut-off values of approximately 0.35, precision increases to over 50 % and below cut-off values of approximately 0.25, recall increases to over 50 %. While the performance at level 1 is good, it decreases significantly with greater tree depths. Still, a cut-off 0.35 at level three, would provide average precision 0.64 and F1 = 0.16. The F1 measure then increases to 0.2 in for similarity cut-off 0.32, and to 0.3 when the cut-off is below 0.26. Overall, the performance at levels 1 and 2 is good, at level three acceptable, and it then decreases significantly for depths greater than three. Here, a cut-off 0.2 at level two, would provide average precision 0.74 and F1 = 0.55; at level three the precision would be 0.6 and F1 = 0.44 (this variation is illustrated by Fig. 5).

3.3.3. Evaluation conclusions

We evaluated several combinations of the three parameters: similarity cut-off [0.15,…, 0.5], MeSH tree depth [1,…,7] and major vs. all MeSH classes. For those combinations, we have calculated the average precision, recall, and F1 measures.

As expected, higher cut-offs yield higher precision results and lower recall results. Moreover, lower depths yield both better precision and recall. Compared to evaluation of major MeSH classes, the precision for this evaluation is much better, while recall results for major MeSH classes are slightly better than for this evaluation. Roughly, for tree depth 3, over cut-off values of approximately 0.25 precision increases to over 70 % and below the same cut-off recall is around 30 %. At tree depth three, which is the aim for news item annotation, results are of acceptable quality considering the aim to give emphasis to precision at level three. In conclusion, classification of all MeSH classes performs significantly better at the desired depth of classification, depth three. At

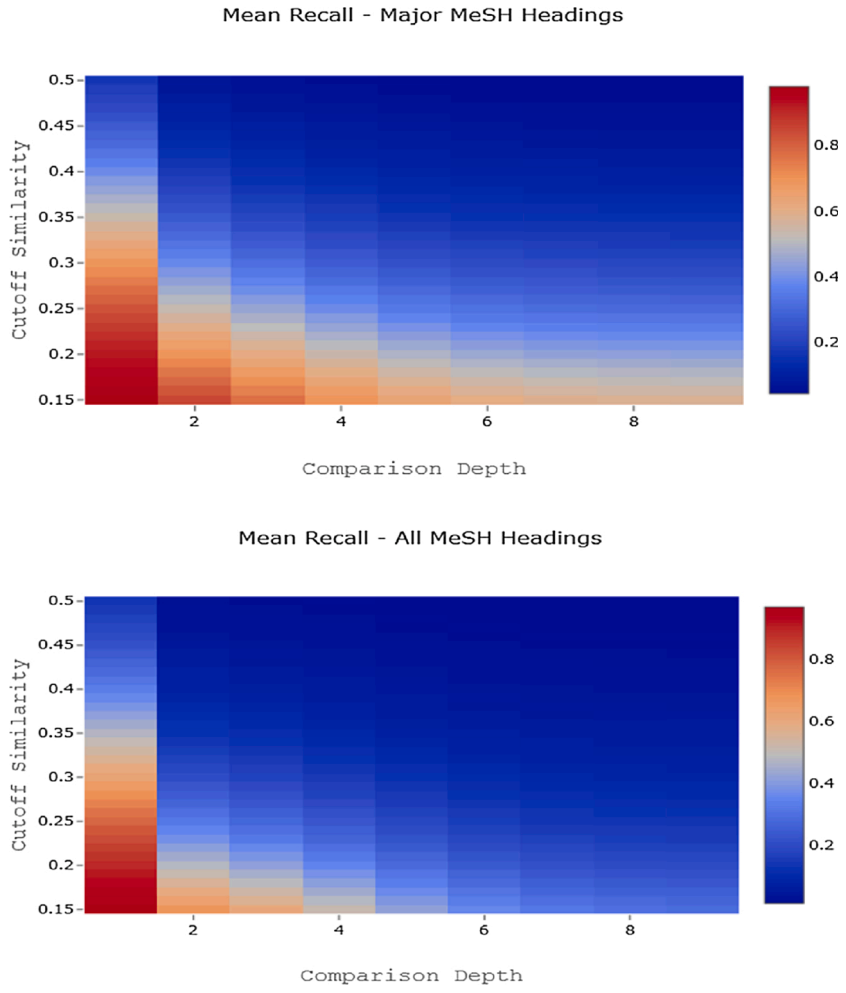


Fig. 4. Precision in the comparison between the MeSH tree depth and the cut-off based on similarity between major MeSH classes (above) and all MeSH classes (below).

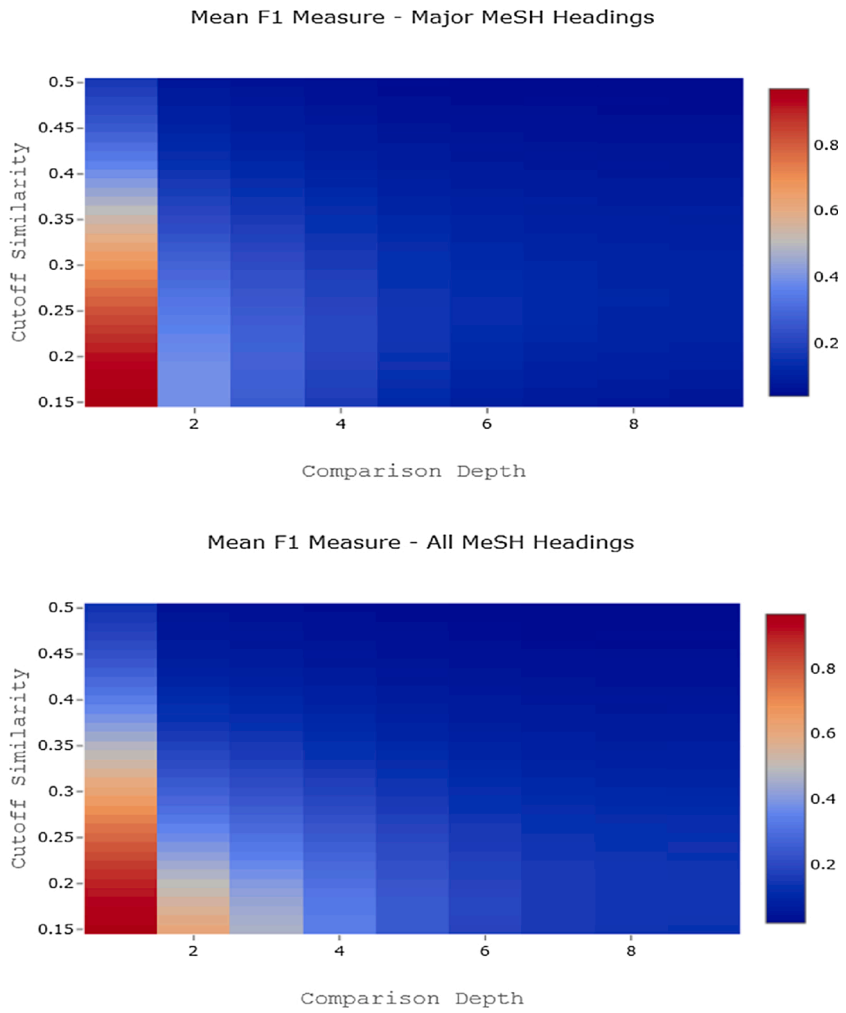


Fig. 5. F1 measures in the comparison between major MeSH classes (above) and all MeSH classes (below), distinguishing the MeSH tree depth and the cut-off based on similarity for major MeSH classes.

that classification depth it is estimated that the optimal similarity cut-off is around 0.2, with precision 0.6 and F1 = 0.44. At that same depth (depth 3), the XLNet based Large Pretrained Transformer classifier achieves a superior F1 = 0.56, a very significant improvement when comparing with the results of the nearest centroid classifier.

3.4. Evaluation results for the annotation of MeSH classes over news articles

3.4.1. Evaluation setting

In a second phase of this study, we evaluated the classifier in the context of news articles. For this purpose, we asked five experts (i.e., health professionals with experience in the usage of MeSH) to annotate news articles using the MeSH classes.

Based on the analysis of the prior evaluation over research articles, we considered that the annotation could go up to a fourth level of deepness in the MeSH tree. Thus, we proceeded with providing each of the five experts with a set of news articles and a spreadsheet where they should annotate with three to ten MeSH classes each of the articles. The appropriate MESH Id can be consulted and obtained at the *NIH MeSH Tree View*[3] and *NIH MeSH Search*[4] . In that spreadsheet, each line was an article and the MeSH classes that are annotating it. An example of this annotation, in the context of diabetes, is the annotation of the news article *"Obesity Stigma And Yo-Yo Dieting Are Behind Chronic Health Conditions, True or False?"* [5] with the following MeSH descriptors: Obesity D009765; Diet, Reducing D004038; Chronic Disease D002908; Statistics as Topic D013223; Social Stigma D057545; Causality D015984; Correlation of Data D000078331; Risk D012306; Complementary Therapies D000529; Epidemiologic Methods D004812.

For MeSH-based manual annotation, we selected as news topics the five health domains that correspond to recent priorities of European public health authorities [6]: mental health, diabetes mellitus, coronavirus, childhood obesity, and child in care. In the following paragraphs, we present the results of the evaluation for the first three health scenarios that we chose to analyse in depth. In the conclusions section, show the full range perspective covering the evaluation of the classifier over the five health topics. This will provide us with a range of different examples, which can lead to some conclusions regarding the annotation of health-related news through the MeSH classifier.

(MC1) The term *mental health* (MeSH ID 68008603) exists in MeSH with the unique ID D008603 since 1967. It is defined as the emotional, psychological, and social well-being of an individual or group. It falls on two paths in the MeSH tree under the roots *Psychiatry and Psychology* MeSH category (at deepness 3) and *Health Care* MeSH category (at deepness 4). There are 81433 scientific articles hand-annotated with this

[3] https://meshb.nlm.nih.gov/treeView
[4] https://meshb.nlm.nih.gov/search
[5] https://www.edgemedianetwork.com/health_fitness/health//282003

MeSH class in the 2019 version of MEDLINE we use.

(MC2) The term *diabetes mellitus* (MeSH ID 68003920) exists in MeSH with the unique ID D003920 since 1984. It is defined as a heterogeneous group of disorders characterized by hyperglycemia and glucose intolerance. It falls on two paths in the MeSH tree under the root *Diseases Category* (at deepness 3 and 5). The 2019 version of MEDLINE we use here includes 315341 scientific articles hand-annotated with this MeSH class.

(MC3) The term coronavirus (MeSH ID 68017934) also exists in the MeSH tree with the unique ID D017934 since 1994. It is defined as being part of the CORONAVIRIDAE disease family, which causes respiratory or gastrointestinal disease in a variety of vertebrates. It falls on a unique path in the MeSH tree under the roots Coronaviridae (at depth 6). There are 5976 scientific articles hand-annotated with this MeSH class in the 2019 version of MEDLINE we use.

3.4.2. Evaluation results

The diagrams in Fig. 6 show the evaluation results where the X-axis is the tree depth, the Y-axis is the similarity cut-off threshold, and the colour code is the result of the evaluation (precision, recall, and F1 measures). Table 3 summarizes the most important results and Table 4 compares the results of both classifiers.

Higher cut-off thresholds yield higher precision results and lower recall results. Moreover, lower depths yield both better precision and recall. Roughly, in the case (MC1) for depth 3, over cut-off values of approximately 0.35 precision increases to over 50 % and below cut-off values of approximately 0.25 recall increases to over 50 %. In the case (MC2), for three depth 3, over cut-off values of approximately 0.27 precision is reaching 80 % and below cut-off values of approximately 0.25 recall decreases to over 60 %. Similarly, in the case (MC3), for three depth 3, over cut-off values of approximately 0.37 precision increases to over 60 % and below cut-off values of approximately 0.23 recall increases to over 50 %.

For the case (MC1), a cut-off around 0.3 at level two would provide F1 above 50 %. The case (MC2), over a cut-off 0.35 at level two, shows F1 above 0.5. Finally, in the case (MC3), a cut-off of 0.35 at level three, would provide average precision 0.64 and F1 = 0.11.

3.4.3. Evaluation conclusions

The study has evaluated combinations of the two parameters - similarity cut-off and MeSH three depth - for each of the five MIDAS use cases. We have calculated the average precision, recall, and F1 measure for each of these. In conclusion, the classification of MeSH classes over news articles performs significantly better at the desired depth of classification, depth three. Though the variation between the evaluation of the classification of news articles with different health topics vary over a small range up to deepness 3 much increasing after that.

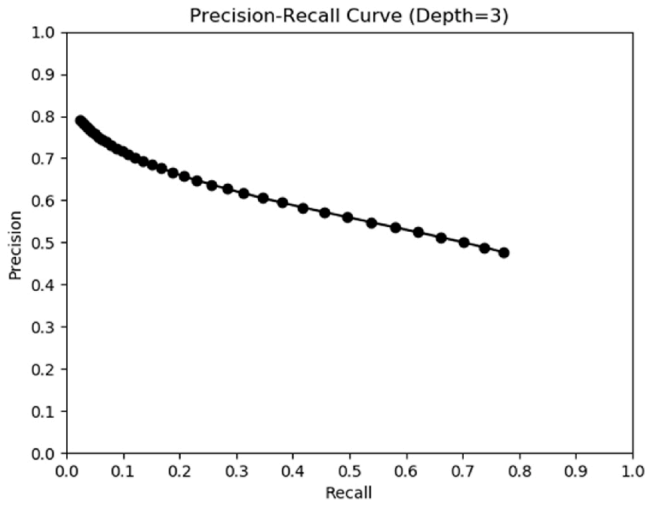


Fig. 6. Precision-recall curve contributing to the optimal choice of the appropriate cut-off at MeSH tree depth 3.

Table 3
The results of the Nearest Centroid MeSH classifier evaluation on news throughout five different health domains, describing the optimal threshold values (cut-off and F1) per MeSH tree depth.

Health Domains	*MeSH Tree Depth*			
	1	2	3	4
	F1	F1	F1	F1
Mental Health	0.930	0.5515	0.2905	0.0839
Diabetes Mellitus	0.9442	0.5066	0.3422	0.1196
Coronavirus	0.8873	0.3448	0.2790	0.0814
Childhood Obesity	1	0.7619	0.6355	0.2998
Children in Care	0.9883	0.5765	0.3315	0.1402

Table 4
Comparison of both MeSH classifiers results (F1-score) on our news dataset at a fixed label depth of 3.

Health Domains	*Classifier*	
	Nearest Centroid	XLNet
Mental Health	0.2905	0.2743
Diabetes Mellitus	0.3422	0.4388
Coronavirus	0.2790	0.3692
Childhood Obesity	0.6355	0.4842
Children in Care	0.3315	0.2345

The results are good up to tree depth three in all cases, where the F1 measure yields similar results. Moreover, the performance at level one is good, and decreases significantly with greater tree depths. At that classification depth it is estimated that the optimal similarity cut-off is around 0.2, providing on average a F1 measure above 40 %. In all the cases analysed, at tree depth three, which is what the researchers aim for the annotation of news articles, the results vary significantly across the health topics in analysis (Fig. 7).

In the Table 3 we present results for the five health domains studied providing the optimal threshold values (cut-off, F1) per a given depth of MeSH tree search. The poor results in the case *Children in Care* are mostly due to the limited number of news items that effectively discuss the topic, making it difficult to have a reasonable expert hand annotation of the related news articles. The case of the annotation of news on the topic of Coronavirus is also poor, mostly due to the challenging task of the annotation of the news article because the dimension of the topic in the media is not only on the health sphere, but affecting the lifestyle of the population, the economy, and many other aspects relevant for society In general.

Similarly, the topic mental health in news articles is poorly annotated. It is represented by a large variety of aspects of the disease in society and the way authors express these in their content. Although this is a topic of growing awareness and exposure, it is still not fully understood and accepted by the general public. This might be a substantial part of the challenge for this specific topic, in the sense that it is sometimes introducing ambiguity when comparing with other topics that are treated in a more precise manner by news outlets (e.g., diabetes).

4. Conclusions and future work

We have proposed an innovative approach to advanced search of health-related news to support the workflow of health professionals. The proposed health classification enables the exploration of news and extraction of health-related insights, based on the MeSH vocabulary. One of the impactful applications of that annotation is the exploration of news events and articles in a similar way as in PubMed, with which most researchers in the health domain are familiar with.

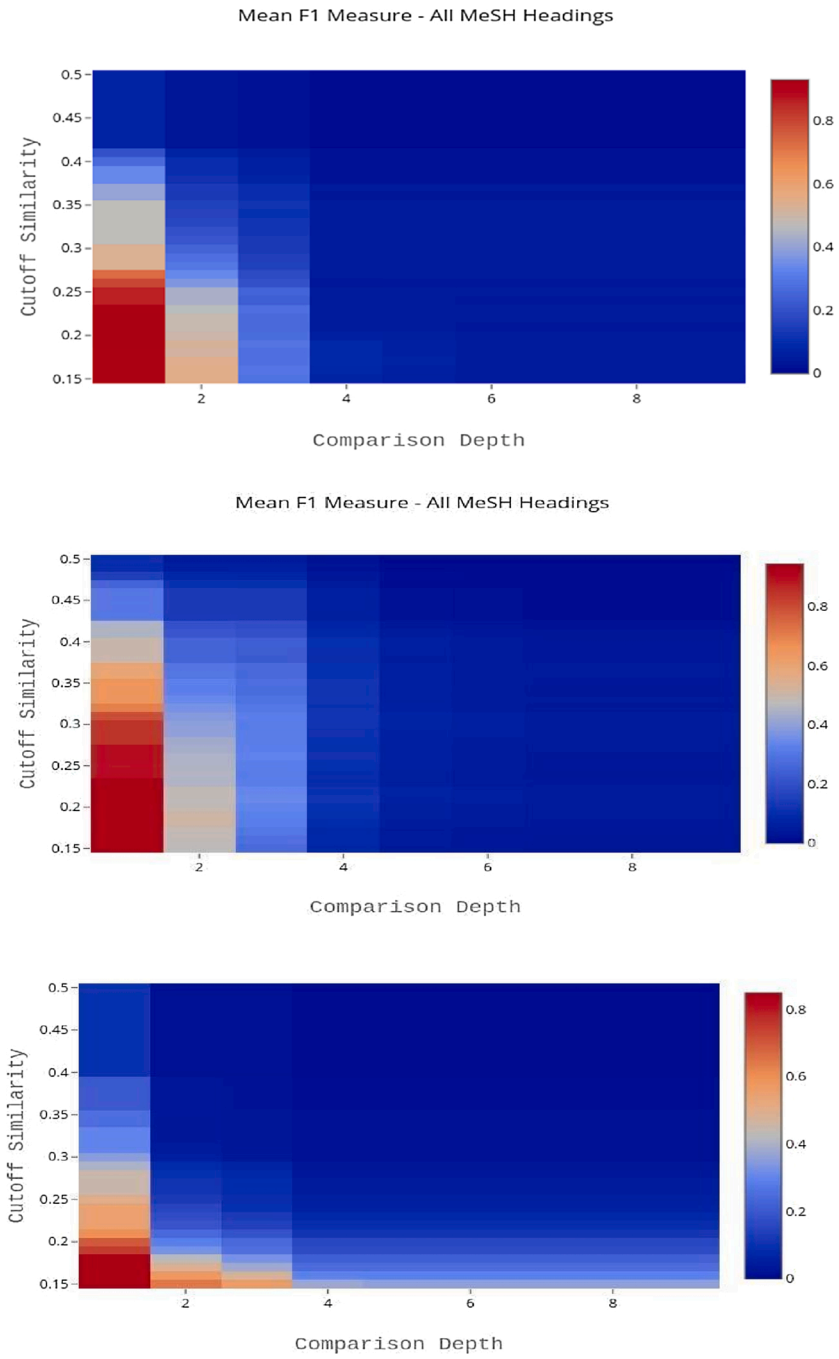


Fig. 7. F1 measure in the comparison between the evaluation of news articles with focus on mental health (above), diabetes mellitus (middle) and Coronavirus (below), considering the MeSH tree depth and the cut-off based on the similarity for major MeSH classes.

The experimental evaluation shows that it is possible to leverage the annotations of scientific articles to automatically annotate news articles. For these, we invited domain experts to annotate the articles with MeSH classes and confronted those with the annotation of our classifier. The classification of health-related news articles is the main contribution of this paper. Text in scientific articles and news is very different, effectively a difference in domain. But there is also a marked difference across different health topics and how they are exposed in the media. One such example is the difference between the coverage of mental health and diabetes, mostly due to the fact that mental health is still a topic that fails to be considered in full, perhaps, due to social stigma or to its perception as a less important topic.

Table 4, which compares both classifiers on the news data, shows that the Nearest Centroid classifier outperformed the XLNet based classifier in 3 out of the 5 health domains. This is surprising given that the XLNet classifier significantly outperforms the Nearest Centroid classifier on the medical papers dataset.

The automated evaluation in this paper considered several combinations of relevant parameters: a similarity cut-off; and the MeSH three depth. In the case of scientific articles, we compared the evaluation using only the major MeSH classes (annotated by domain experts) against all MeSH classes. In the case of news articles, we compared the evaluation of news relating to five public health policy domains: mental health, diabetes, coronavirus, childhood obesity, and children in care.

For the above combinations, we have calculated the average precision, recall, and F1 measures. With this analysis we conclude that the

classification of scientific articles with all MeSH classes performs significantly better at the MeSH tree depth of classification 3. At that classification depth it is estimated that, for the Nearest Centroid classifier, the optimal similarity cut-off is around 0.2, with an average F1 measure around 0.4. In the case of diabetes, the three depth 4 over cut-off values of approximately 0.38 show precision increase to over 60 %, while below cut-off values of approximately 0.36 recall increases to over 51 %. In the classification of news articles, the health domain that it relates to and the frequency of news seems to have an impact in the evaluation. This study shows that news articles about diabetes get better evaluation results than those on the Coronavirus mostly because of the diversity in the scope of news reflecting the impact in several domains other than health, and the gap to that learned over scientific articles. Predictably, classifying health news at a depth greater than 3 is challenging. The additional specificity of labels easily leads to increases in sporadic correlations between news texts and scientific articles that a text classifier can pick up, leading to a decrease in precision. Simultaneously, even when details can allow a human expert to label at those greater depths, this knowledge is hard to capture given a classifier trained exclusively on scientific articles.

Taking into consideration the results obtained in the classification of health-related news, we will further explore the novelty presented by this MeSH classifier in that domain that entails its own challenges. Future work includes the improvements to the classifier itself, through a differentiated assignment of importance to the MeSH tree branches, refining those that are taken into consideration, and using weights to distinguish the relevance of the classes. Another envisaged improvement is the appropriate inclusion of the information obtained by the qualifiers (that, unlike the descriptors used as MeSH classes in the proposed classifier, provide complementary information). With regards to the Large Pretrained Transformer classifier, results should improve with domain adaptation. BioBERT [15] leveraged unsupervised pretraining to improve results in scientific article classification. BERTMeSH [42], showed further improvements via supervised pretraining (fine-tuning) on the specific dataset to be classified. Both unsupervised and supervised pretraining on health news data are possible. Unsupervised pretraining only requires the categorization of news articles as being health-related which is not difficult. Supervised pretraining would require clever tricks to match news articles to labeled scientific literature or a costly amount of expert hand labeling. An exploration of other domain adaptation techniques (between scientific articles and news articles) could also help push the results of on news further. Another avenue that should be explored in terms of news article classification is in identifying labels to prune: i.e. labels that may make sense in scientific literature but are unlikely to be meaningful in health news. How to perform such pruning automatically or semi-automatically is an open question.

Further research also includes the evaluation of news articles on a wider range of health-related topics (including, e.g., asthma) which will require a substantial increase of domain experts to annotate with the MeSH classes a selection of related articles. This will provide us with a larger perspective on the efficiency of the classifier across the public health and healthcare scope. It will also help us better understand where learning from the MEDLINE dataset (and from the narrower scope scientific articles it includes) is sufficient to have a satisfactory result on the classification of news articles related to those health topics.

The specific challenges in the hand annotation of MEDLINE articles (where one can annotate with a term and the reader can assume that related terms are represented) might impact the efficiency of the built classifier, which is learning from this labelled dataset. The precision to which the MeSH tree can reach in many health domains is reflected in the choice of MESH classes that are used by the MEDLINE experts to classify these scientific articles. We suspect that the human assumption both from the side of the expert providing the hand annotation and the human reader, make equivalent the annotation of two slightly different MeSH classes. The automated classification does not make these assumptions based on the multitude of parameters inherent to the context of the scientific article, but humans can. This can be partially solved by relying on higher nodes in the tree. Though, an automated classification that aims to provide MeSH classes deeper in the tree would need to tackle this problem.

Furthermore, the evaluation parameters obtained will be used to further optimise the classifier and evaluate the classifier improving its classification of news articles. The proposed classifier enables the user to follow a workflow similar to that of exploring scientific articles in *PubMed* when monitoring health news, and to extract further insights from the monitored news previously annotated (e.g., using the MeSH headings in their search, as proposed in [27]). It also enables new functionalities that are based on the MeSH terminology (see Fig. 1 for an example of a data visualization module allowing the user to account for the percentage of news articles that talk about a specific MeSH heading related to the search topic).

We aim to further explore the integration of this MeSH text classifier with the exploration and monitoring of local and worldwide news, as well as in social media. This is an important task in Public Health, impacted by the choice of the appropriate parameters that can express the defined priorities. The accurate monitoring of worldwide news contributes to a global perspective of world health, but also to the aspects of regional health where public health institutes can act. Though, this will require building it from a cross-lingual classifier. It can also contribute to the evaluation of the success of public health campaigns by allowing decision-makers to assess what the news media's response was to them, often reflecting the opinion of their communities. Moreover, the further exploration of health news articles can help health professionals to avoid news bias in the era of *fake news* [25].

Declaration of Competing Interest

There are no conflicts of interest

Acknowledgment

This work was supported by the European Commission H2020 project MIDAS (G.A. nr. 727721).

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.artmed.2021.102053.

References

[1] Aronson A, et al. The NLM indexing initiative's medical text indexer, vol. 89. Medinfo; 2004.
[2] ArcGis. WHO's COVID-19 disease monitoring. url: https://experience.arcgis.com/experience/685d0ace521648f8a5beeeee1b9125cd. Accessed: 20 March 2020. 2020.
[3] Baker S, Korhonen AL. Initializing neural networks for hierarchical multi-label text classification. Association for Computational Linguistics. BioNLP 2017, Association for Computational Linguistics. 2017. p. 307–15.
[4] Black M, et al. Meaningful integration of data, analytics and services of computer-based medical systems: the MIDAS touch. 32nd IEEE CBMS International Symposium on Computer-Based Medical Systems 2019.
[5] Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004;32(suppl_1):D267–70.
[6] Boilson A, Connolly R, Staines A, Davis P, Connolly J, Weston D. Improving European Healthcare Systems through the Development of a Realist Evaluation Framework for a European Public Health Data Analytic Project. Biomed Central (BMC) Implement Sci J. 2019.
[7] Dai S, You R, Lu Z, Huang X, Mamitsuka H, Zhu S. FullMeSH: improving large-scale MeSH indexing with full text. Bioinformatics 2020;36(5):1533–41.
[8] Devlin J, Chang M-W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1; 2019. p. 4171–86 (Long and Short Papers).
[9] Fang L, Karakiulakis G, Roth M. Are patients with hypertension and diabetes mellitus at increased risk for COVID-19 infection? Lancet Respir Med 2020. https://doi.org/10.1016/S2213-2600(20)30116-8.

[10] Henderson L, Lachlan. Automated text classification in the DMOZ hierarchy. TR; 2009.
[11] Humphrey SM. MedIndEx system: medical indexing expert system. Inf Process Manag 1989;25(1):73–88.
[12] Jin Q, Dhingra B, Cohen W, Lu X. AttentionMeSH: simple, effective and interpretable automatic MeSH indexer. Proceedings of the 6th BioASQ Workshop A Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering 2018:47–56.
[13] Kaggle. COVID-19 open research dataset challenge - CORD-19. url: https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge. Accessed: 20 March 2020. 2020.
[14] Kilicoglu H, et al. Semantic MEDLINE: a web application for managing the results of PubMed searches. Proceedings of the Third International Symposium for Semantic Mining in Biomedicine, Vol. 2008; 2008. p. 69–76.
[15] Lee J, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36(4):1234–40.
[16] Lindberg DA. Internet access to the national library of medicine. Effect Clin Pract: ECP 2000;3(5):256.
[17] Liu K, Peng S, Wu J, Zhai C, Mamitsuka H, Zhu S. MeSHLabeler: improving the accuracy of large-scale MeSH indexing by integrating diverse evidence. Bioinformatics 2015;31(12):i339–47.
[18] Manning C, et al. Introduction to information retrieval. Cambridge Univ. Press; 2008.
[19] Mao Y, Zhiyong L. MeSH Now: automatic MeSH indexing at PubMed scale via learning to rank. J Biomed Semantics 2017;8(1):15.
[20] Midas Project. A MIDAS contribution to the global COVID-19 monitoring strategy. url: http://www.midasproject.eu/2020/03/13/a-midas-contribution-to-the-global-covid-19-strategy/. Accessed: 20 March 2020. 2020.
[21] Mladenic D. Turning Yahoo into an automatic web-page classifier. In: Prade H, editor. Proceedings of European Conference on Artificial Intelligence (ECAI). Chichester [etc.]; 1998. p. 473–4. John Wiley & Sons.
[22] Mladenic D, Grobelnik M. Feature selection on hierarchy of web documents. Decis Support Syst 2003;35:45–87.
[23] Newman D, Karimi S, Cavedon L. Using topic models to interpret MEDLINE's medical subject headings. Australasian Joint Conference on Artificial Intelligence 2009:270–9. Springer, Berlin, Heidelberg, 2009.
[24] Peng S, You R, Wang H, Zhai C, Mamitsuka H, Zhu S. DeepMeSH: deep semantic representation for improving large-scale MeSH indexing. Bioinformatics 2016;32(12):i70–9.
[25] Pita Costa J, et al. Health News bias and epidemic intelligence for public health. Proceedings of the SIKDD 2019 2019.
[26] Pita Costa J, et al. MIDAS hand-annotated news articles [dataset]. Zenodo; 2020. https://doi.org/10.5281/zenodo.4034281.
[27] Pita Costa J, et al. Text mining open datasets to support public health. Conf. Proceedings of WITS 2017 2017.
[28] Pita Costa J, et al. The meaningfulness of open data in public health and healthcare. Proceedings of the 12th European Public Health Conference 2019 2019.
[29] Rankin D, et al. The MIDAS platform: facilitating the utilisation of healthcare Big data in Northern Ireland and beyond. the 8th Annual Translational Medicine Conference. Clinical Translational Research and Innov. Centre (C-TRIC) 2017.
[30] Rogers FB. Medical subject headings. Bull Med Libr Assoc 1963;51(1):114–6.
[31] Sarrouti M, Alaoui Said Ouatik El. SemBioNLQA: a semantic biomedical question answering system for retrieving exact and ideal answers to natural language questions. Artif Intell Med 2020;102:101767.
[32] Srinivasan P, Libbus B. Mining MEDLINE for implicit links between dietary substances and diseases. Bioinformatics 2004;20(Suppl 1):i290–6.
[33] Yan Y, et al. Biomedical literature classification with a CNNs-based hybrid learning network. PLoS One 2018;13(7):e0197933.
[34] UNESCO International Research Institute on Artificial Intelligence – IRCAI. IRCAI's COVID-19 disease monitoring. url: http://coronaviruswatch.ircai.org/. Accessed in: 20 March 2020. 2020.
[35] U.S. National Library of Medicine – NLM. MEDLINE dataset. url: https://www.nlm.nih.gov/databases/download/pubmed_medline.html. Accessed in: 1 September 2020. 2020.
[36] U.S. National Library of Medicine. MeSH data. url: https://www.nlm.nih.gov/databases/download/mesh.html. Accessed in: 1 September 2020. 2020.
[37] U.S. National Library of Medicine. MeSH on demand. url: https://meshb.nlm.nih.gov/MeSHonDemand. Accessed: 20 March 2020. 2020.
[38] Wang A, Singh A, Michael J, Hill F, Levy O, Bowman SR. Glue: a multi-task benchmark and analysis platform for natural language understanding}. 7th International Conference on Learning Representations, ICLR 2019 2019.
[39] World Health Organisation – WHO. WHO Director-General's opening remarks at the media briefing on COVID-19—11. 11 March 2020. url: https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19—11-march-2020. Accessed: 20 March 2020. 2020.
[40] Xiang Yu-Tao, Yang Yuan, Li Wen, Zhang Ling, Zhang Qinge, Cheung Teris, et al. Timely mental health care for the 2019 novel coronavirus outbreak is urgently needed. Lancet Psychiatry 2020;7(3):228–9.
[41] Xun G, Jha K, Yuan Y, Wang Y, Zhang A. MeSHProbeNet: a self-attentive probe net for MeSH indexing. Bioinformatics 2019;35(19):3794–802.
[42] You R, Liu Y, Mamitsuka H, Zhu S. BERTMeSH: deep contextual representation learning for large-scale high-performance MeSH indexing with full text. bioRxiv 2020.

Chapter 3

Multimodal Learning

This chapter includes the work published in the article titled *Multimodal metadata assignment for cultural heritage artifacts* (Rei et al., 2023). The article was published in the journal Multimedia Systems, Volume 29, issue 2, dated April 2023, in November 2022.

We developed a multimodal classifier for inferring metadata of digitized silk fabric artifacts using a late fusion approach. The three modalities are Image, Text, and Tabular data. We created our multilingual and multitask text classifier using XLM-RoBERTa (Conneau et al., 2020) which was pre-trained on web text in 100 different languages. This development was informed by our initial exploration of the problem in Rei and Mladenić (2020). The image classifier uses a ResNet (He et al., 2016) pre-trained on ImageNet. The development of this classifier was informed by Dorozynski et al. (2019). Tabular data and late fusion used Xgboost (T. Chen & Guestrin, 2016). The labels were provided by the final version of the SILKNOW Thesaurus (Alba et al., 2022; Léon et al., 2019) with the classification targets being based on a higher level of the tree than in Rei and Mladenić (2020). Unlike Pita Costa et al. (2021) where the depth of the tree was determined automatically, the determination of the depth in this work was guided by experts and the need to have a minimum number of examples across modalities. The categories of metadata inferred in this work are the same as in Rei and Mladenić (2020) under slightly different names: technique, timespan, material and place. The text descriptions are written in English, Spanish, French, or Catalan. Of the individual modalities, when evaluating examples where the modality exists, the text classifier achieved the highest test average F1 score at 86%, followed by the image classifier at 58%, then the tabular classifier at 56%. Instead, when evaluating all examples, and considering each missing modality example as an error for that modality, the best individual results correspond to the tabular classifier with an average F1 score of 56%, followed by the image classifier with 52%, and finally the text classifier at 49% average F1. The difference is that while 97% of digitized artifacts have an associated image, only 39% have a corresponding text description. The best result is obtained by the Multimodal with an average F1 score of 74%, significantly higher than any individual modality classifier. In conclusion, we showed that we were in fact able to accurately predict missing properties in the digitized silk fabric artifacts. While the quality of predictions varied between individual modalities, we showed that the multimodal approach provided the best practical results.

Multimedia Systems
https://doi.org/10.1007/s00530-022-01025-2

REGULAR PAPER

Multimodal metadata assignment for cultural heritage artifacts

Luis Rei[1,2] · Dunja Mladenic[1] · Mareike Dorozynski[3] · Franz Rottensteiner[3] · Thomas Schleider[4] · Raphaël Troncy[4] · Jorge Sebastián Lozano[5] · Mar Gaitán Salvatella[5]

Received: 30 May 2022 / Accepted: 11 November 2022

Abstract
We develop a multimodal classifier for the cultural heritage domain using a late fusion approach and introduce a novel dataset. The three modalities are Image, Text, and Tabular data. We based the image classifier on a ResNet convolutional neural network architecture and the text classifier on a multilingual transformer architecture (XML-Roberta). Both are trained as multitask classifiers. Tabular data and late fusion are handled by Gradient Tree Boosting. We also show how we leveraged a specific data model and taxonomy in a Knowledge Graph to create the dataset and to store classification results.

Keywords Cultural heritage · Multimodal · Deep learning · Text classification · Multilingual · Image classification · Transformer · Convolutional neural networks

1 Introduction

1.1 Motivation

Some domains within cultural heritage domains deal with knowledge that is not broadly known by the public, but only by domain experts. Despite many objects having been digitized, even those experts still struggle to find in online catalogs what they are looking for. Thus, they are forced to return to the cumbersome manual consultation of published catalogs, or even card files. If such is the situation for experts, the broader public is still more removed from access to that information. The European production of silk fabrics is an example of one such domain. It is witness to an essential field of European and global history, linked to world trade routes, the production of luxury goods of enormous symbolic importance, technological developments and the very advent of the Industrial Revolution. However, the material vulnerability of these objects and the institutional fragility of many local heritage organizations has rendered it relatively hidden to the public. As regards the information about that heritage, many descriptions, and images of objects exist within in-house databases that are only available as local files. In other cases, those records are uploaded by many museums across the globe, in siloed repositories and heterogeneous, often incompatible formats. A few of

✉ Luis Rei
luis.rei@ijs.si

Dunja Mladenic
dunja.mladenic@ijs.si

Mareike Dorozynski
dorozynski@ipi.uni-hannover.de

Franz Rottensteiner
rottensteiner@ipi.uni-hannover.de

Thomas Schleider
thomas.schleider@eurecom.fr

Raphaël Troncy
raphael.troncy@eurecom.fr

Jorge Sebastián Lozano
jorge.sebastian@uv.es

Mar Gaitán Salvatella
m.gaisal@uv.es

[1] Department for Artificial Intelligence, Jošef Stefan Institute, Jamova Cesta 39, 1000 Ljubljana, Slovenia

[2] Jošef Stefan Institute International Postgraduate School, Jamova Cesta 39, 1000 Ljubljana, Slovenia

[3] Institute of Photogrammetry and GeoInformation, Leibniz University Hannover, Nienburger Straße 1, D-30167 Hannover, Germany

[4] EURECOM, Campus SophiaTech, 450 Route des Chappes, 06410 Biot, France

[5] Departamento de Historia del Arte, Universitat de València, Spain, Av. Blasco Ibáñez 28, 46010 Valencia, Spain

Published online: 21 November 2022

them give public access to the images and metadata of such silk objects through APIs, many more through their websites, but harmonization and integration efforts have been very scarce. Therefore, it is very hard for general audiences, historical experts, and industry (e.g., fashion designers) to access this knowledge.

One possible application for a cultural heritage domain such as European silk fabrics are Exploratory search engines, which help users to explore a topic of interest [46]. They enable serendipitous discovery of items, and they are especially appropriate when these items come with rich structured metadata. ADASilk [1], named after Ada Lovelace, is such an exploratory search engine, based on a knowledge graph (KG), that enables to search and browse silk fabrics objects for both domain experts and users not necessarily familiar with this topic. Thus, not only historians or scholars, but also designers or simply fans of fashion can access such a significant and little-known part of our heritage.

Some records have essential information, like the production year or the weaving techniques, semantically annotated, others include it only in rich textual descriptions, and for some objects it is not available at all. These missing metadata can be considered as gaps that potentially could be filled in. Thanks to the progress in natural language processing, information extraction, and image processing, there are now techniques that can help to address such problems. Digitization of culturally significant assets is a time-consuming process that requires experts and funding. This often forces a cultural institution to make a trade-off between the number of objects digitized and the effort per object. Less effort per object often implies a smaller number of details captured, less strict guidelines, and sometimes mistakes. Nevertheless, this area could benefit from automated aids for collection caretakers, that often must catalog similar or identical objects scattered across the world. Obtaining predictions or suggestions for their description and possible matching pieces would be a great help for that task, taking also into account the many objects still waiting to be properly cataloged.

This paper presents methods that enable further annotation of these museum objects through a multimodal classification approach that trains models to predict such missing metadata from images, text descriptions and other (available) metadata. The outcome is then further used to enrich an underlying knowledge graph that feeds the ADASilk exploratory search engine. Domain experts can easily assess the quality of the automatically generated annotations through rich visualization and connections between the items.

1.2 Hypothesis

Our first hypothesis is that we can predict, fairly accurately, a set of domain-relevant properties of cultural heritage objects (silk fabrics) from images and text descriptions. Our second hypothesis is that a multimodal approach involving both images, text descriptions and additional knowledge about other properties than those to be predicted will produce better results than any method relying on a single modality. In this context, the term "better" refers to both, the quality of the results and the number of objects for which this information is inferred. That is, we expect the multimodal approach to result in more correct predictions and in predictions for a larger number of objects than the single modality methods. These hypotheses will be evaluated in the context of digitized metadata of silk fabric artifacts with data originating in multiple museums.

1.3 Contributions

The main scientific contributions of this paper are related to our research hypotheses. We introduce a multimodal machine learning approach, adapted to the cultural heritage domain, for predicting properties of digitized artifacts. We perform an in-depth analysis of the performance of our classification models, i.e., models based on individual modalities and the multimodal classifier. Additionally, we introduce a novel dataset[2] to the cultural heritage and multimodal analysis domains that includes data for four different tasks and three different modalities. It consists of harmonized text and image data from heterogeneous, multilingual sources that went through different stages of preprocessing, cleaning, and enrichment like domain expert-guided entity linking and grouping.

Finally, we show how our metadata predictions can be properly represented through classes and properties in our data model, which includes using information about a.o. their time stamp and or the used algorithm, and consequently integrated into existing Knowledge Graphs.

1.4 Challenges

The challenges faced in this work can be split broadly into those pertaining to the creation of the dataset and those related to the automated annotation. The latter ones can be further categorized according to the modality that is used for predicting the properties of the objects.

[1] https://ada.silknow.org/.

[2] https://zenodo.org/record/6590957.

1.4.1 Data and labels

The data used in this work belongs to the cultural heritage domain. More specifically, it is related to silk textiles produced in Europe, primarily in the period between the 15th and the 19th centuries. In the domain of cultural heritage, we cannot expect all class labels to be equally likely or equally correlated. For example, in some locations, more silk fabric objects were produced than in others. Similarly, we know that the production of silk fabric objects in a given location likely started after a certain point in time and possibly subsided after a certain date. We also know that catalogs are curated by humans and often have strong thematic biases. For example, certain museums focus almost exclusively on objects created within one location. The data we use in this work was aggregated from different sources. That is, it was crawled from 12 different museum or collection websites. Each museum may have different standards for how it collected the underlying objects and how it digitized the information related to these objects. Importantly, this gives each museum its own standards for how to write text descriptions, how to create images, and how to annotate properties. Regarding these properties of digitized artifacts, accurately representing them requires adequate data modelling capabilities and considerable domain expert collaboration. This collaboration is also important in creating a dataset for machine learning. Labels need to be mapped from annotations made in different languages and grouped into domain relevant classes. Due to the partially automated nature required to create the dataset, challenges arise that are common in such processes: label text requires normalization such as correcting typos, unifying the styles of dates, and matching different locations to specific countries. Errors made in this process can often be systematic, for example, a failure to link a specific value of a property due to the form of writing it particular to that catalog will likely result in that value not being present in all records originating in that catalog.

1.4.2 Image classification

In the context of this paper, the classification of images aims to predict abstract properties of the silk fabrics depicted in the images. Whereas it may be relatively straightforward to learn to classify the material of a depicted piece of fabric, the prediction of semantic information such as the production place of the fabric, the period of time in which the fabric was produced or the technique used to manufacture the fabric is assumed to be much more challenging. Furthermore, it is assumed that there are interdependencies between the these properties of silk fabrics, e.g. a certain production technique may only have been used in a certain period of time. This is why multi-task learning is investigated for image classification. However, standard multi-task classification frameworks require one reference label for every task to be learned during training for every training sample; The challenge we have to face is that in real world data, as they were collected for the dataset presented in this paper, there may be many training samples for which annotations are unavailable for some of the target variables to be predicted. Accordingly, this fact must be taken into account in the training of a multi-task classifier. Additionally, the available number of class labels constituting the class distribution of a variable is often imbalanced for real-world datasets. This constitutes a further challenge to supervised learning, which is addressed by utilizing a suitable training strategy for the image classification method.

1.4.3 Text classification

Supervised approaches are often challenging to perform with data from the cultural heritage domain for several reasons. Text descriptions are not present for the majority of objects in an archive. Many of the text descriptions that are available, in most museums, tend to be short sentences, almost title-like. In specific domains, such as the cultural heritage of silk production, many of the terms used in the text are very domain specific. Each museum has their own standard of how and what to write in a text description: some may focus on the history of the objects and write very grammatical paragraph-length descriptions meant to be read by the public, others may focus on the properties of the object and write a single enumerating sentence, and others still, may focus solely on the depictions or visual patterns of an object. Finally, museums are spread geographically, and thus we can expect to deal with multiple languages, making our problem multilingual and cross-lingual. To summarize, we end up with a small collection of domain specific texts, written in different languages, with different content both semantically and syntactically, and wildly varying lengths. These texts are then associated with labels, based on the provided properties of the object. As already discussed, these labels are not all equally likely or correlated, and many of these accidental regularities are likely to interact with the language and the particularities of the text style of the museum.

1.4.4 Multimodal classification

One of the challenges in this work is that we want to integrate predictions made from images and text. Most work done in the literature, is exclusive to depictions or type of object: the image shows a scene or object and the text describes it. In our case, there may be no scene depicted in an object, and we do not consider describing the object beyond certain properties. For example, if we have a fabric that shows a certain pattern, describing the visual shapes of the pattern (e.g., triangles) is not a goal. Rather, we

need to deduce, from the image, properties of how, when, where, and with what the object was made. Similarly, with text descriptions, there may be a good amount of words that describe visual patterns, scenes depicted, and historical facts associated with the object, but the goal is, again, to determine those same intrinsic properties of the object's making. Another challenge that is uncommon is the reduced and variable overlap between images and text descriptions. Not only is our work subject to a comparatively small dataset, restricted by historical reality and difficulties of data collection, but we must also deal with the fact that for most archives of culturally relevant objects, many objects that have been photographed have no corresponding textual description. In fact, we'll see that less than half of all objects have both these modalities. Another challenge, uncommon outside of retrieval scenarios, is that we can have multiple different images, with different angles and focus, per each individual object while it makes no sense to talk about multiple text descriptions per object. Yet another challenge we need to deal with, common to many real world applications but not to research datasets, is that we do not have all properties for all objects. For example, for a given object, we might know what material and techniques were used but not when or where it was made. Finally, our dataset, although drawn from several museums, contains under 30k objects and approximately 11k text descriptions. Effectively making it small compared to general datasets, but not uncommonly so for a dataset in the cultural heritage domain.

2 Related work

2.1 Cultural heritage domain

Since the development of the web, many Cultural Heritage (CH) organizations provide metadata on their items through some search engine, APIs or aggregators. Unfortunately, there has been little unity in the data formats, which makes data integration a complex task. One solution to this problem is in the case of museum data is the use of Semantic Web technology and more specifically the development of Knowledge Graphs based on ontologies that follow the open CIDOC Conceptual Reference Model (CRM). CIDOC-CRM was developed for this purpose, i.e., to facilitate inter-museum data integration. It provides many relevant classes and properties to represent domain-specific CH objects and is easily extendable. It is the outcome of more than 20 years of development by ICOM's International Committee for Documentation (CIDOC) [19]. CIDOC-CRM can, however, only be considered a starting point for ontologies that deal with museum data, such as in our case. The fact that it can be easily extended makes it, however, easy to add more domain-specific classes and properties as they are needed in projects such as ours.

There are more and more efforts of different CH organizations to integrate their data with Semantic Web technology and building knowledge graphs: CultureSampo is the result of integrating heterogeneous cultural content [28]. The challenges consisted amongst others of converting legacy data into linked data and making it heterogeneous. Getty ULAN was used as structured vocabulary to find connections between two referenced persons, for example. One similar example is ArchOnto [34], which specifically addresses the challenges of CH data from and for national archives. Both can be an inspiration for work such as ours in this paper, but given how fine-grained the vocabularies in Cultural Heritage domains, such as ours, can be, it is still necessary to deal with the languages and domain specific vocabulary differently in each case.

The training data used for the experiment in this paper is fully extracted from the SILKNOW Knowledge Graph that relies on classes and properties defined by CIDOC-CRM and its direct extensions CRMsci (Scientific Observation Model) and CRMdig (Model for provenance metadata). All our data is therefore part of the specific CH domain of "silk fabrics" and accordingly semantically annotated and enriched, for example through linking and normalization of properties, such as used materials and weaving techniques.

2.2 Knowledge graphs and culture AI

Knowledge graphs allow the representation of multi-source information about many entities and their relationships to each other. The data stored in a Knowledge graph can then be used for many other tasks, especially when structured knowledge of a specific domain is relevant, e.g. the development of product designs [37]. Other common domain-specific fields are Medicine, Cyber Security and Finance, but Knowledge graphs are also used a lot to aid product development and research for language-based tasks such as question answering systems, recommender systems and information retrieval [23, 64]. A knowledge graph can also help with textual metadata-aided visual pattern extraction and recognition [11], which is very relevant for this paper. Lastly, as we deal not only with images, but also textual metadata, the SemArt project can be considered related: it is a multi-modal dataset for semantic art understanding. Unlike in this study, they did, however, not work towards metadata completion, but focused only on retrieval [24] .

2.3 Image classification

Applying and adapting machine learning techniques to support solving tasks in the context of preserving the cultural heritage is a growing field of research. Many works address

image-based classification of artworks by training an image classifier on the basis of images with known class labels to make predictions for images with unknown properties [21]. First works investigate classical machine learning approaches aiming to predict characteristics of a depicted painting [3]. In [7], one-versus-all Support Vector Machines are trained based on HOG features (histograms of oriented gradients, [16]) of images showing paintings, with the goal to predict the artist of the painting.

Instead of training a classifier to predict variables based on handcrafted image features, Convolutional Neural Networks (CNNs) allow for simultaneously learning features from given input images as well as learning a mapping of these features to class scores based on labeled training images [35, 36]. Thus, a trained CNN can be used to predict a class label for an object with unknown properties from an image depicting that object. CNN-based classifiers are also applied in many works addressing attribute prediction for depicted objects in the context of cultural heritage, where the focus is on making predictions for images showing paintings [11, 52]. In [58], the *artist*, the *genre* as well as the *style* of a painting are learned by means of a variant of AlexNet [35], achieving on average 68.3% correctly classified images for the three variables using the WikiArt dataset (http://www.wikiart.org/). Investigating the prediction of a painting's *artist*, Sur and Blaine [57] obtain 82.5% overall accuracy on the Rijksmuseum dataset [44] utilizing a ResNet18 [26]. In both cases, there is one CNN per classification task, and network weights pre-trained on a variant of the ImageNet dataset [17] are used to improve the classification performance.

Instead of training a separate CNN per task to be learned, the concept of multitask learning aims to exploit interdependencies between related tasks by means of jointly learning them in one network and, thus, to improve the network's performance in solving the individual tasks [10]. Multi-task learning for CNNs is addressed in many recent works [15] investigating different strategies for combining the training of several tasks. In the domain of cultural heritage, the most frequently used strategy applies a feature extraction network producing a high-level image representation that is shared among all tasks and which is processed by additional task-specific layers designed to solve the individual classification tasks, e.g., [5, 25, 56]. These works do not only perform multitask learning for predicting characteristics of paintings on the basis of images, they also make use of pre-trained CNNs for the shared feature extraction network.

In contrast to all works cited so far, which are dedicated to the classification of paintings, we address the CNN-based classification of images of silk fabrics. Even though there are papers dealing with the CNN-based classification of images of textiles, e.g. [29, 43, 48, 61], distinguishing different textile patterns, no work could be found addressing the classification of images of fabrics in the context of cultural heritage except for our previous one. The image classification network presented in this paper can be seen as an expansion of [20], aiming to predict different properties of silk fabrics; the network takes images of silk fabrics as input, where a high-level image representation produced by a fine-tuned ResNet [27] is shared among all task-specific classification branches that deliver the predictions. In contrast to [20] as well as [5, 25, 56], we adapt the training of the network weights such that hard training examples get a higher impact on the weight updates. In this way we want to deal with the problem of class imbalance in the training data, aiming to improve the classification performance for underrepresented classes. For that purpose, we combine a variant of the focal loss [38] with the multi-task softmax cross entropy loss used in [20], leading to a training strategy that focuses on hard training examples in a multi-task scenario while allowing for missing class labels at training time for some of the tasks to be learned. Furthermore, we investigate the prediction of four variables instead of three like in [20] and evaluate our methodology on a much larger dataset consisting of images from several museum collections instead of one collection only.

2.4 Text classification

Much of the recent work in natural language processing has focused on fine-tuning large transformer neural networks [59] pretrained as language models such as BERT [18] and RoBERTA [41]. The most common approach is to add a task-specific head to the pretrained transformer to create the final model architecture. The full model is then trained on the task specific data. This is the process that is called fine-tuning. On most natural language processing (NLP) tasks, some variation of this approach provides the best results. Previously, many multilingual and cross-lingual artificial intelligence approaches to text used pretrained and aligned word-embeddings, such as the aligned fastText embeddings [8, 30]. But similarly to the overall trend in NLP, recent approaches have also moved towards using fine-tuning of pretrained transformers. Our work follows this trend, we fine-tune the pretrained XLM-R [14]. Multitask models are often very desirable from a practical perspective: a single model is easier to deploy and maintain, offers faster inference and occupies less space in memory when compared to multiple models. Further, multitask learning can often result in measurable improvements [9]. [40] showed that multitask training of BERT improved results across several tasks.

The use of Natural Language Processing, especially text classification, in the cultural heritage domain isn't very widespread. This is a consequence of the fact that the digitization of artifacts usually includes images and some labels or tags, but text descriptions are far less common. The highlight is the work of [51] on text descriptions of paintings

from the Rijksmuseum Amsterdam. They used an Information Extraction approach rather than classification. Their pipeline included Named Entity Recognition, Part-of-Speech tagging and dependency parsing to extract concepts from the text, those concepts were then matched to an ontology and finally classified according to a role. These roles included all the properties we use: Technique, Material, Date, Place, plus others such as "Creator" of the artwork, style, and the subject depicted. Their data, although limited to a total of 250 text descriptions, was manually annotated and each text contained a concept-role pairing. They reported an average $F1$ of 61.2% compared to non-expert human average of 65.1%. Our work differs from this in several key areas. First, our classification approach is more generalizable, as it does not necessarily require information to be directly present in the text and more resilient to misspellings and non-standard grammar. In fact, even correctly linking information known to represent a certain label (e.g. material) from tables can be challenging in the presence of spelling issues. Second, we work with multilingual data from multiple sources, which presents additional challenges. Finally, our dataset contains many more samples and uses automatic labelling based on information present in catalogs rather than externally annotated.

2.5 Tabular classification

Gradient Boosted Decision Trees (GBDT) [22] has long been the state of the art and the common choice for handling tabular data. Concurrently with our work, Neural Network based alternatives have been proposed which can outperform GBDT in certain situations [2, 31]. To the best of our knowledge, there has been no work published on tabular classification in the cultural heritage domain that we can provide an overview of.

2.6 Multimodal classification in cultural heritage

[4] presented a joint image-text neural network architecture for classifying images of paintings by artist and year. Their text input consisted of a limited set of labels (style, media, and genre) rather than text descriptions. Conceptually, this is similar to CLIP-Art [13], an application of CLIP [49] to the retrieval of artwork images. CLIP learns to associate a small text vocabulary akin to labels with images through joint contrastive pre-training. This was applied to The iMet Collection dataset [63], possibly the most similar dataset to our own, it includes images of artworks associated with labels (also called "tags") that describe what is visually depicted in the object (e.g. "Dragons"), its visible properties (dimensions, medium) as well as other culturally relevant properties (e.g., country of origin). Another very relevant dataset is in this context is Artpedia [55], a dataset of images of paintings associated with textual descriptions tagged as either "visual sentences" that describe the scene depicted in the painting or as "contextual sentences" that describe other aspects of the painting such as its historical context. The tasks for which this dataset was created consist of separating visual from contextual sentences and the retrieval of the correct image for a given text. The differences between the related work and our work are clear. We propose to handle images, multilingual text, and tabular data as equal modalities. We also propose to handle data from multiple collections.

3 Data

3.1 Knowledge graph

The SILKNOW Knowledge Graph[3] lies at the center of all efforts to create a unified representation of the metadata of European silk textiles, particularly from the 15th to the 19th century. All the data used in our experiments was downloaded from 16 sources, most of them are public online museum records, for which we built crawling and harvesting software. In addition to that, we have data from the SILKNOW[4] project partners Garin and the University of Palermo (Sicily Cultural Heritage). The dataset used in the experiments was created from a full export of all objects in the knowledge graph, which consists of the metadata of 38,873 unique silk objects before any preprocessing steps. This export includes in total 74,527 unique image files.

To model this heterogeneous data from so many sources, we chose and relied strongly on the CIDOC Conceptual Reference Model (CRM). We also developed our own SILKNOW ontology[5] to extend CIDOC-CRM with further classes and properties for cases where it did not cover some specifics of the silk textile domain and also for some extra information. For example, the confidence score for metadata predictions, once we started integrating the results of those predictions back to the KG.

To develop a converter[6] that could unify all the original data with all these classes and properties into one knowledge base, mappings have been created by domain experts. And on a technical level, all museum records had to be harvested and were first converted into a common JSON file format through our crawler software[7] but each array inside this format still had the original field labels from the museums before the final conversion. For example: the majority of

[3] https://zenodo.org/record/5743090.
[4] https://silknow.eu/.
[5] https://ontome.net/namespace/36.
[6] https://github.com/silknow/converter/.
[7] https://github.com/silknow/crawler/.

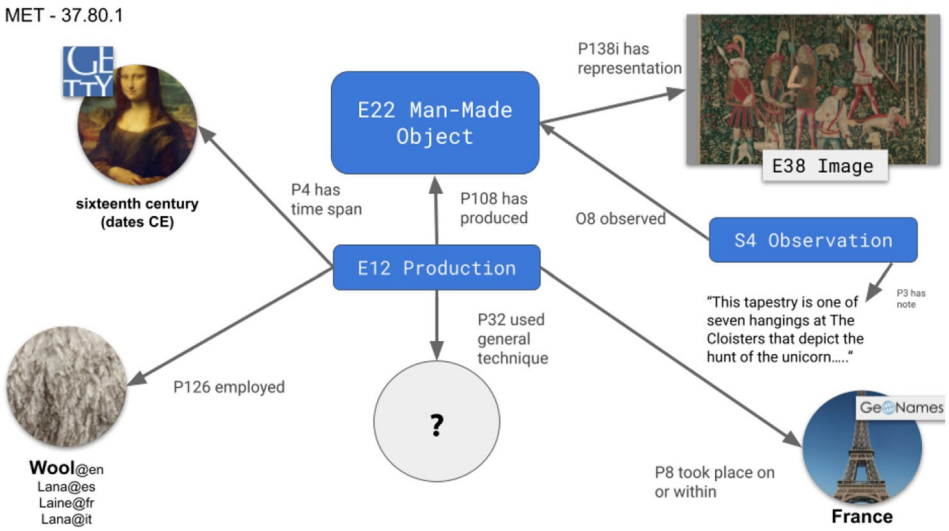


Fig. 1 A record from the MET museum with a missing property represented in the knowledge graph using our ontology and controlled vocabularies

museums have a field for describing the production time of a silk object, but in most cases museums use different names for their field. Moreover, the museums are from all over the world and we are facing different languages for both the field names and their values. This is why we created a mapping for, e.g, a field named "Date" (Metropolitan Museum of Arts) and the class `E12_Production` with the property `P4_has_time-span` and another class `E52_Time-Span`. Likewise, a mapping rule will be written for the field named "date_text" (API of the Victoria and Albert Museum) and for the (Spanish) field named "Datación" (Red Digital de Colecciones de Museos de España).

Another very central part of our knowledge representation is the SILKNOW Thesaurus[8], a controlled vocabulary which contains many explicit and multilingual concept definitions for materials, techniques, and motif depictions relevant for these silk textiles. Thanks to this thesaurus, a lot of information and entities from very explicit categorical fields of the original museum records could be linked, without any advanced machine learning techniques - the string literal could just be matched with the (multilingual) labels of the thesaurus and then replaced with a unique concept link. This explicit representation of knowledge forms the core of the dataset used to predict missing metadata. This includes cases where a categorical value is either not given at all or "hidden" in longer textual descriptions and not explicitly semantically annotated.

Once all the modelling, download, conversion and enrichment steps were taken, the final knowledge graph was uploaded onto a SPARQL endpoint from where all the data across languages and museums can be queried the same way. To make access easier, we also developed a RESTful API, so it is not necessary for web developers to write SPARQL queries, and an aforementioned exploratory search engine on top of this API, called ADASilk. It is aimed at users with only little technical background or little background knowledge about the domain of silk, to make them able to discover a lot of the data in the KG. ADASilk offers an advanced search with many filters, some topic suggestions, and in general a clean visual interface that shows all objects with their images and metadata.

3.2 Extracting and normalizing labels

The development of the SILKNOW Knowledge Graph is a combined effort of data processing that relies on a data modelling and annotation process created in collaboration with domain experts. This is especially true for the SILKNOW Thesaurus. The group labels used in the experiments in this paper are based on the hierarchy and relations of concepts of the silk textile domain described in this controlled vocabulary. As described in Sect. 3.1, a big part of categorical property values could be easily extracted, linked and through the string replacement indirectly automatically normalized thanks to the SILKNOW Thesaurus. This means that many concepts are accessible even though there were originally different strings, including typos in some cases, synonyms, or translations. An example would be a weaving technique like "Damask", which would be "Damas" in French and "Damasco" in Spanish and Italian: for all these, we replace the string literal with one link to the same concept. In addition to the SILKNOW Thesaurus, we also use linked open data like such as GeoNames[9] to normalize and link place names.

[8] https://skosmos.silknow.org/thesaurus/.

[9] https://www.geonames.org/.

Matching strings with such thesaurus or other controlled vocabularies was not without challenges. As will be also explained in more detail in Sect. 6, misspellings or unique punctuation could still cause the matching process not to work properly. To give an example: If the string value of a record was "silk; gold thread" the latter would not have been linked, due to a bug that did not properly consider a semicolon as a separator. Other such cases existed as well, as the development of the SILKNOW Knowledge Graph is an ongoing process and concurrent with this work. See Fig. 1 for an illustration of a museum record in the knowledge graph.

The aforementioned hierarchy defined in the SILKNOW Thesaurus can be used to select specific types or subtypes of properties. To refer back to the previous example, we could select only objects with the weaving technique "Damask", but also only objects made with "Two-coloured damask" which is even more specific. Based on the Thesaurus, we can also make sure that we only choose objects based on equivalent levels of this hierarchy.

Based on these enrichments and the linking process, we created a pipeline to extract the dataset based on pre-specified criteria. We first developed a comprehensive SPARQL query that outputs all museum objects described in the Knowledge Graph (KG) and includes, if available, the most relevant properties: the identifier of the object in the knowledge graph, the museum where the description comes from, the text description, and URL links to the images that illustrate the object. The results of this query were exported as a CSV file, which we then post-processed to make sure that we have a format of one row per object. In this final format, the CSV is used as the basis for all experiments.

3.3 Label grouping

In principle, the Knowledge Graph contents can be used to generate training and test samples for the classifiers described in Sect. 4. One would just have to associate the images and/or the text given for a record with the annotations in the categorical variables of interest. The available annotations can be easily converted into class labels. However, a statistical analysis of these annotations revealed that most of them occur very rarely in the data, while for all categorical variables there were one or a few classes which were dominant in the sense that many records belonged to them. Supervised classifiers have problems with imbalanced training data sets, and it would seem very difficult for a classifier to successfully differentiate classes for which it has seen only a very small number of training samples, if on the other hand there are thousands of samples for some other classes. To still be able to extract meaningful information from the available modalities using supervised methods while at the same time having the chance of achieving a reasonably good classification performance, a simplified class structure was defined. Domain experts analyzed the class distributions and aggregated classes corresponding to different categories into compound classes. Care was taken for the aggregated classes to be consistent with the Thesaurus, and aggregated only if they were considered to be related according to the domain experts. At the same time, the aggregation was guided by the frequency of occurrence of class labels so that the compound classes would occur frequently enough to be used for training the supervised classifiers described in Sect. 4.

The resultant simplified class structure was integrated into the Knowledge Graph in the form of so-called *group* fields, which were made available for all semantic properties of interest, principally, the ones corresponding to the different tasks in this work. Such grouping was applied to the following properties: Material, Technique, production place (with a country granularity), production time (with the century granularity) and the object type or object domain group, to be able to filter out non-textiles that use silk. Grouping was not an easy task, domain experts had to deal with more than 200 concepts that had to be grouped according to the aforementioned categories. Techniques were the most complex to group. To do so, domain experts grouped the concepts according to two fundamental criteria: (1) whether they belonged to the same hierarchy, for example, velvet and its types. In fact, there are many types of velvet, classified depending on the nature of the pile such as broderie velvet, ciselè velvet, cut velvet, pile-on-pile velvet, uncut velvet, etc. (2) If they were somehow related to a certain technique, for example, the effects obtained of applying differently warp and weft, that is, whenever a yarn is introduced into a fabric to produce an effect or pattern. On the other hand, materials were not complex as they were made in large groups according to their origin, that means according to the product obtained from the processing of one or more raw materials, in the course of which their structure has been chemically modified, e.g. animal fibres are distinguished from vegetable fibres. Using a conversion table for aggregation prepared by the domain experts, the contents of the *group* fields could be derived automatically from the original semantic annotations. Having thus expanded the Knowledge Graph, training, and test samples could be easily generated from it by appropriate SPARQL queries that would export the contents of the *group* fields associated with each record.

3.4 Dataset preparation and properties

The goal of the dataset preparation is the conversion of the knowledge graph data with normalized and grouped labels described, respectively, in Sects. 3.1, 3.2, and 3.3, into a dataset for the experiments in Sect. 6 using the classification methods described in Sect. 4.

Table 1 Class structure and class distribution of the records

Variable name	Class name	Total	Training	Validation	Test
Timespan	19th century	5849	3492	1180	1177
	18th century	4397	2576	901	920
	20th century	2483	1520	483	480
	17th century	1134	689	231	214
	16th century	880	542	180	158
Place	FR	5265	3156	1037	1072
	IT	3205	1853	687	665
	GB	2837	1721	562	554
	ES	2630	1605	521	504
	IN	1190	735	231	224
	CN	699	426	127	146
	IR	671	409	142	120
	JP	533	325	92	116
	TR	331	205	57	69
Technique	Embroidery	3123	1814	657	652
	Velvet	2193	1273	454	466
	Damask	1685	1004	333	348
	Other technique	1150	722	219	209
Material	Animal fibre	17,382	10,387	3445	3550
	Vegetal fibre	2051	1255	396	400
	Metal thread	2046	1223	422	401

Table 2 Names of the museums contributing to the dataset with their identifiers (ID) used in this paper, and distribution of the 28,077 records over the museums for the training (train.), validation (val.) and test sets

Museum name	ID	Total	train.	val.	Test
Metropolitan Museum of Arts	met	6524	3835	1325	1364
CDMT Terrassa	imatex	6119	3690	1204	1225
Victoria and Albert Museum	vam	5527	3300	1133	1094
Rhode Island school of design	risd	3226	1913	634	679
Boston Museum of fine arts	mfa	2610	1579	517	514
Garín 1820	garin	1558	972	300	286
Collection du Mobilier National	mobilier	1293	796	267	241
Red Digital de Colecciones de Museos de España	cer	781	490	142	149
Joconde Database of French Museum collections	joconde	375	224	78	73
Smithsonian Museum	smithsonian	38	29	14	14
Versailles	versailles	18	8	5	5
Art Institute of Chicago	artic	8	4	2	2

The first step was to select the records in the knowledge graph that were relevant to the domain.

The second step as to select only records that contained a value for one of the variables to be predicted, i.e., labeled samples. Uncommon labels, with a total frequency below 150, were discarded. The final step was to randomly split the records into disjoint sets:

- A training set consisting of 60% of the data for supervised learning;
- A validation (or development) set, consisting of 20% of the data for hyperparameter tuning and multimodal supervised learning;
- A test set, consisting of 20% of the data, for evaluation of the proposed method.

Given that the objective is to train and evaluate a multimodal multitask approach on records, that regularities exist within each collection (i.e., museum) that comprises the data, that the text modality is also multilingual, and that both modalities and task specific labels may be missing from a record, we believe the most reasonable way to split the dataset is a random split of records. The distribution of the data per each set and class label can be seen in Table 1.

Table 3 Modality statistics of all records in the dataset that provide a class label for at least one of the variables

Dataset	Total	With image	With text	With image and text	Without images and text
train.	16,840	16,260	6717	6495	358
val.	5602	5419	2184	2101	100
Test	5635	5441	2133	2068	129
Total	28,077	27,120	11,034	10,664	587
	100.0%	96.6%	39.3%	38.0%	2.1%

The values are given for the training (train.), validation (val.) and test sets as well as for the total dataset

The distribution of samples over the museums can be found in Table 2 and an overview of the modalities can be found in Table 3. We can see how 27,120 or 96.60% of the 28,077 records about annotated fabric objects contain at least one image, but only 11,034 or 39.29% of them contains a text description. The overlap consists of 10,664 or 37.98%. The proportion between training validation and test sets in each case corresponds roughly to the aforementioned 60-20-20 split.

Text data in our dataset consists of descriptions of fabrics or objects made mostly of fabrics. These descriptions range in length from a short sentences to multi-sentence

Table 4 Examples of text descriptions present in our dataset

Text description
White and silver striped fabric with supplementary weft of flat silver strips whose floats form vertical stripes with leaves at intervals. White floats of the weft form outlines for serpentine floral sprays spread over the striped areas.
Furnishing fabric, woven, British, c. 1895, Alexander Morton & Co., red/brown plain silk weave
Dibujo Palma en color azul grisáceo Urdimbre: Trama: 36 pasadas Rapport: 65 cm ancho y 104 cm alto (incompleto)

Table 5 Text length in characters and space delimited tokens

	Min	Q1	Median	Mean	Q3	95th percentile	Max
Characters	60	173	343	693	856	2367	16,333
Tokens	7	28	56	115	142	392	2826

Table 6 Language distribution of text descriptions based on language of the museum

	English	Spanish	French	Catalan
Records	7271	1975	1126	680

paragraphs to multi-paragraph texts with thousands of words. Some descriptions focus primarily on a single aspect, such as a scene depicted or the history of the object, while others focus on various properties of the object. Table 4 shows some examples of these descriptions. To eliminate some errors present in the data, we removed any text descriptions smaller than 60 characters. The resulting distribution of lengths is summarized in Table 5. These descriptions are in 4 different languages: English, Spanish, French, and Catalan. The counts for each are shown in Table 6.

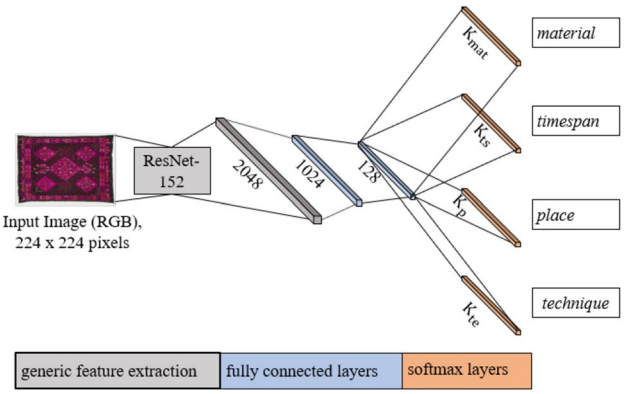


Fig. 2 Network architecture of the CNN for multitask image classification. The input image scaled to 224 x 224 pixels is presented to a pre-trained ResNet-152 (grey) to extract generic features. The resulting 2048-dimensional feature vector is mapped to a domain-specific joint representation of 128 dimensions by two fully connected layers (blue). The task-specific classification branches consist of one softmax layer each (orange) that delivers the class scores for the corresponding variable. K_{mat}, K_{ts}, K_p, and K_{te} denote the number of class labels for the tasks *material*, *timespan*, *place*, and *technique*, respectively

4 Methods

4.1 Image classification

The goal of the image classification is to predict one class label per classification task, i.e., the prediction of a class label for each of the target variables *technique*, *timespan*, *material* and *place*, for an image that illustrates an object. For that purpose, an image classifier is trained using all images of all records contributing to the dataset described in Sect. 3.4. We propose to use a convolutional neural networks (CNN) for that purpose, motivated by the success of CNN in image classification. As there are many records with annotations for more than one of these variables, we propose to train the classifier to predict all classes simultaneously in a multitask framework, exploiting the inherent relations between the variables to learn a joint representation that is used by task-specific classification heads. A detailed description of the chosen network architectures can be found in Sect. 4.1.1, whereas the strategies used for training are presented in Sect. 4.1.2.

4.1.1 Network architecture

Figure 2 shows the structure of the CNN for multitask learning for the prediction of the four target variables. Its input consists of an RGB image scaled to a size of 224 x 224 pixels. This image is presented to the ResNet 152 network of [27] pre-trained on ImageNet [17], which serves as a generic feature extractor for the image [53] and produces a feature vector of 2048 dimensions. We apply dropout with a probability of 10% after this layer [54]. This is followed by $L_{fc} = 2$ fully connected layers, the first one having 1024 and the second one having 128 nodes, which are shared by all tasks.

Rectified linear units [45] are used as nonlinearities in both of these joint layers. They produce a joint representation of the image of $N_r = 128$ dimensions. This representation is processed by four task-specific classification branches, each consisting of one additional softmax layer only, which delivers the class scores $y_{km}(\mathbf{x}, \mathbf{w})$ for the input image $\mathbf{x}$ to belong to class k for variable m. The number of nodes of the softmax layer corresponds to the number of classes to be differentiated for a specific task. The CNN architecture is shown in Fig. 2.

The CNN predicts one class label per task for every image. In case of multiple images per record, one such class label is predicted for each one of the images and the prediction with the highest softmax score is chosen to be the prediction for the record.

4.1.2 Training

In training, the parameters $\mathbf{w}$ of the CNN described in Sect. 4.1.1 are learned by minimizing a loss function $E(\mathbf{w})$. The parameters of our network consist of the parameters $\mathbf{w}_R$ of ResNet-152, which are initialized from a pre-trained model published by [27], and the parameters $\mathbf{w}_{FC}$ of the fully connected and softmax layers, which are initialized randomly by a variant of the Xavier initialization also described in [27]. In the training procedure, we determine the parameters $\mathbf{w}_{Rt}$ of the last NL_{RT} layers of ResNet-152 considering exclusively entire residual blocks and the parameters $\mathbf{w}_{FC}$ of the fully connected layers, whereas the parameters $\mathbf{w}_{Rf}$ of the first $152 - NL_{RT}$ ResNet-152 layers are frozen [62]. Thus, the parameter vector consists of three subsets: $\mathbf{w} = \left(\mathbf{w}_{Rf}^T, \mathbf{w}_{Rt}^T, \mathbf{w}_{FC}^T\right)^T$. NL_{RT} is a hyperparameter to be tuned.

Two loss functions can be used for training the network. The first one, originally proposed in [20], is an extension of the standard softmax cross-entropy loss with weight decay [6]:

$$E_{SCE}(\mathbf{w}) = -\sum_{n=1}^{N}\left(\sum_{m\in M_n}\sum_{k=1}^{K_m} t_{nmk}\cdot ln\left(y_{km}\left(\mathbf{x}_n, \mathbf{w}\right)\right)\right) + \omega_R \cdot R\left(\mathbf{w}_{Rt}, \mathbf{w}_{FC}\right) \tag{1}$$

In Eq. 1, $y_{km}\left(\mathbf{x}_n, \mathbf{w}\right)$ is the softmax score for the nth training image $\mathbf{x}_n$ to belong to class k for variable m. The indicator variable t_{nmk} is one if the class label of sample n for variable m is k and zero otherwise. The sum is taken over all N training samples and K_m classes for task m. M_n is the set of tasks for which the true class label is known for the training sample n, so that the loss in Eq. 1 considers exclusively samples x_n with $t_{nmk} = 1$ for learning task m. In this way, the fact that the annotations for most samples are incomplete, i.e. that annotations are only available for a subset of the variables to be predicted, can be considered. If multiple annotations are available, the corresponding classification losses will be backpropagated to the joint layers from multiple classification branches, thus supporting the learning of a joint representation for all variables. The outputs for variables for which the true class label is unknown will not contribute to the loss and to the parameter update. Finally, the term $R\left(\mathbf{w}_{Rt}, \mathbf{w}_{FC}\right)$ corresponds to regularization by weight decay, which is only applied to the parameters to be updated in training; ω_R is a hyperparameter defining the influence of this term on the result.

One problem of the data described in Sect. 3.4 is its imbalanced class distribution. In this case, minimizing the cross-entropy loss in Eq. 1 will favor the dominant classes, resulting in a poor performance for the underrepresented ones. To mitigate these problems, a multi-class extension of the focal loss [38, 39] with regularization is utilized for training:

$$E_F(\mathbf{w}) = -\sum_{m\in M_n}\left(\sum_{n=1}^{N}\sum_{k=1}^{K_m}\left(1 - y_{km}\left(x_n, \mathbf{w}\right)\right)^{\gamma}\cdot t_{nmk}\cdot ln\left(y_{km}\left(x_n, \mathbf{w}\right)\right)\right) + \omega_R \cdot R\left(\mathbf{w}_{Rt}, \mathbf{w}_{FC}\right) \tag{2}$$

The only difference between the loss functions in Eqs. 1 and 2 is the penalty term $\left(1 - y_{km}\left(x_n, \mathbf{w}\right)\right)^{\gamma}$, where γ is a hyperparameter modulating the influence of this term on the result. This penalty term forces the loss to put more emphasis on samples that are difficult to classify (having a small score y_{km} for the correct class). Assuming the samples of underrepresented classes to be hard to classify by the CNN, this loss is expected to improve the results for these classes.

Starting from initial values derived in the way described earlier, stochastic minibatch gradient descent based on the ADAM optimizer [32] is applied to determine the CNN parameters, using the default parameters ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$) and a minibatch size of 300. The base learning rate η is another hyperparameter to be tuned. We use early stopping and use the model parameters leading to the lowest loss on the validation set.

4.2 Text classification

Our problem is defined as value prediction for certain properties of an object, a silk fabric, given its text description, which can be written in any one of the four languages listed in Table 6. We have 4 tasks, each denominated according to the property of the underlying fabric object we want to predict: the *technique* and *material* used to create it, the *timespan* or time period when it created, and the *place* where it was created. While some descriptions directly contain some of this information, as seen in Table 4, this is sufficiently uncommon to prevent a purely extractive approach

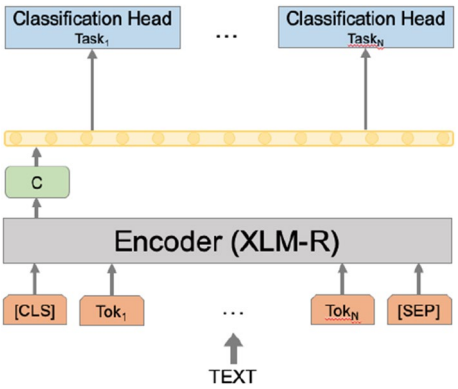


Fig. 3 Multitask architecture: a shared XLM-R based encoder followed by task specific classification heads. The input to each classification head is the ouput of the transformer "C" corresponding to the input token "[CLS]"

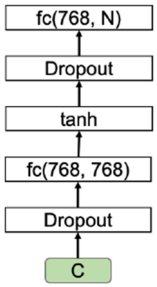


Fig. 4 Task specific classification head: a fully connected (FC) layer followed by a tanh activation, followed by the output projection FC layer. Dropout is applied before both FC layers

from yielding good results. For example, of the 3 texts we showed, only one gives any indication as to where it was produced (*"British"*). We instead rely on regularities present in the text descriptions to make informed guesses. More technically, we frame our problem as a multiclass, multitask, multilingual text classification problem. That is, given a text description of a fabric, written in any language, we want to assign exactly one label out of a set of mutually exclusive class labels for each of the properties we wish to predict, i.e., the tasks.

The text classifier uses a hard parameter sharing based multitask architecture [50], shown in Fig. 3. It consists of a shared encoder followed by task-specific classification heads. The encoder is the multilingual large pretrained transformer, XML-R [14]. Following the method outlined in [18], a special classification token, CLS, is prepended to all inputs. The final hidden state corresponding to this token is used as the aggregate sequence representation. It is the only transformer output forwarded to the classification heads. All classification heads are identical except for the output dimension of the last layer, the output projection layer, which equals the number of classes of the task. A diagram of a classification head is shown in Fig. 4. A softmax function can convert the output logits of the last layer to normalized probabilities.

To finetune our transformer-based classifier, at each step, a task is randomly selected using proportional sampling. A batch of examples for this task is then created and fed to the classifier. The cross entropy loss is then calculated and weights adjusted through backpropagation. Adam [33] is used as the optimizer with weight decay [42].

4.3 Tabular classification

When considering Knowledge Graph records of objects, we can represent them as structured data. That is, a table where each row represents an object and each column a property. We use four separate task-specific classifiers to perform tabular classification. These all use the same learning algorithm, Gradient Boosted Decision Trees (GBDT) [22], implemented in XGBoost [12]. The input to the tabular classifier consists of the categorical values of non-target variables plus the identifier for the museum, as shown in Table 7. We replace missing values for a feature with a predefined value, represented by the symbol "[NA]" ("Not Available") in the table. The output of the classifier, for each example, consists of N-dimensional logit vector. It is used with the softmax function to predict a target class out of N possible classes. This classifier trained by gradient descent to minimize the cross-entropy loss.

4.3.1 Hyperparameters

While a detailed explanation of each hyperparameter that control the resulting model and learning of GBTs is beyond the scope of this work, we believe some contextualization

Table 7 Tabular Classification, one example input row per task

Target variable	Target value	Feature				
		Museum	Place	Timpespan	Technique	Material
Place	FR	risd	–	[NA]	[NA]	Animal fibre
Timespan	XVIII	met	[NA]	–	Embroidery	Animal fibre
Technique	Other technique	garin	ES	XX	–	Vegetal fibre
Material	Vegetable fibre	vam	GB	XIX	Embroidery	–

Note: time label format changed to roman numbers for ease of readability

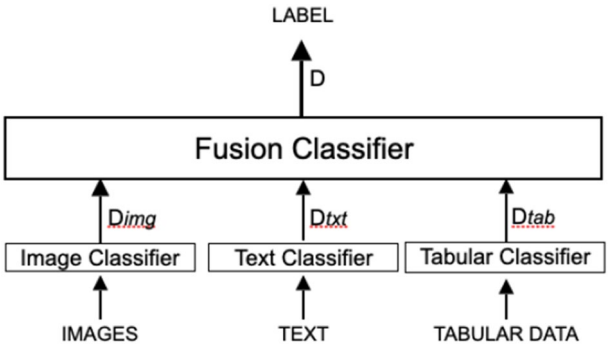


Fig. 5 Architecture of the multimodal classifier. Each classifier based on a single modality takes its own independent decision, D_c, which serves as input to the multimodal classifier. The final decision D is taken by the multimodal classifier, predicting a task-specific label and assigning it to the record

is required. This is due to the relatively larger number of hyperparameters tuned for GBTs in Sect. 5.3 compared to the Neural Network based methods used for the other modalities, and for the convenience of the reader.

The hyperparameters *max_depth* (maximum depth of a tree), *min_child_weight* (minimum weight for tree partitioning), and *gamma* (minimum loss reduction for tree partition) all directly control model complexity, which in turn can have significant consequences in terms of fitting. The hyperparameters *subsample* (the percentage of data sampled per iteration) and *colsample_bytree* (the ratio of features sampled per iteration) can reduce overfitting by adding random noise to the iterative tree building process. Finally, the *learning rate* and *number of rounds* control, respectively, the amount of learning per round and the total amount of learning (i.e., the total number of trees).

4.4 Multimodal classification

Our approach to multimodal classification, shown in Fig. 5, follows a decision level late fusion approach, in which the decision (prediction) from each of the 3 modalities serves as the input to a classifier that takes the final decision on which label to assign to the record. We choose the GBDT algorithm for the multimodal classifier. The input is just one column for each of the three modalities, each column containing the class labels predicted by the corresponding classifier for all of the tasks. If a modality is missing, the values in the corresponding column are set to the missing value indicator [NA], just in the way missing class labels are considered by the tabular classifier. Thus, the multimodal classifier can cope with incomplete records (i.e. records with missing modalities) by design. We created a separate multimodal classifier for each task, i.e. no multitask learning is applied in multimodal classification. .

Table 8 Hyperparameters tuned (image classification)

Hyperparameter	Range	Best
Learning rate η	[1e-5, 1e-3]	1e-4
Weight decay ω_R	[0.0, 1e-5]	1e-3
Degree of fine-tuning NL_{RT}	[0, 36]	30 (E_F) 15 (E_{SCE})
Loss $E(\mathbf{w})$	$\{E_{SCE}(\mathbf{w})$ (Eq. 1), $E_F(\mathbf{w})$ (Eq. 2)$\}$	Focal

An optimal variant is obtained with η=1e-4, ω_R=1e-3, NL_{RT}=30 (i.e., 10 residual blocks), with the focal loss $E_F(\mathbf{w})$

There are several advantages to late fusion over early or intermediate level fusion in our case. Firstly, each record may have multiple images but a single text description. Effectively, the input dimensionality is different. With late fusion, we allow the image classifier to deal with it independently, e.g., by classifying multiple images for the same object and picking the decision with the highest confidence. Secondly, the decisions, represented by a one-hot class vector, have a smaller dimensionality than intermediate representations and thus are more appropriate for scenarios with few samples, which is a common problem in the context of our domain (cultural heritage).

5 Experiments and results

5.1 Image classification

For all experiments in the frame of image classification, we use the split of the dataset described in Sect. 3.4 in order to train the CNN for image classification presented in Sect. 4.1.1 by means of the training strategy described in Sect. 4.1.2. We use all images that are assigned to a record for training and classification, assigning the class labels of the corresponding records to all images associated with it. As pointed out in Sect. 4.1.1, for records associated with multiple images, all images are classified by the CNN at test time, and the image-based prediction having the highest class score is chosen to be the final result.

Experimental setup The workflow of our experiments is as follows: The training dataset is used to update the weights $\left(\mathbf{w}_{Rt}^T, \mathbf{w}_{FC}^T\right)^T$ of the CNN with early stopping. The model parametrization and hyperparameters leading to the lowest loss are calculated on the validation set. In this context, we tuned the hyperparameters listed in Table 8 and described in Sect. 4.1, choosing the values achieving the highest average $F1$ scores on the validation set. Table 8 also presents the selected hyperparameter values. Finally, all test set records for which at least one image is available are used for an

Table 9 $F1$ scores ($F1$) and overall accuracies (OA) of the image classifier obtained by minimizing the Softmax loss (Eq. 1) and the focal loss (Eq. 2) both for the validation and the test sets (evaluated per record)

	Variable	Validation set		Test set	
		$F1$ [%]	OA [%]	$F1$ [%]	OA [%]
Focal loss	Place	49.2	62.5	47.0	63.1
	Timespan	58.4	63.8	57.5	64.5
	Technique	75.5	79.0	77.9	80.2
	Material	52.2	80.6	51.2	80.6
	Average	58.8	71.5	58.4	72.1
Softmax loss	Place	48.2	61.0	47.2	62.2
	Timespan	56.0	64.4	54.2	64.9
	Technique	72.2	75.8	74.0	76.8
	Material	45.0	79.4	43.4	80.7
	Average	55.4	70.2	54.7	71.2
Δ	Average	3.4	1.3	3.7	0.9

Δ gives the difference between the quality metrics achieved using the focal loss and the softmax loss

independent evaluation, using the hyperparameters values tuned on the validation set.

We will report the overall accuracies as well as the average $F1$ scores of the best CNN variant in terms of the average $F1$ score obtained on the test and validation sets for two variants: the first CNN variant is trained by minimizing the softmax cross-entropy loss (Eq. 1), whereas the second variant is trained by minimizing the focal loss (Eq. 2). The overall accuracy *OA* describes the percentage of correctly classified images, denoted as true positives *TP*, among all classified images. As the *OA* is biased towards classes with more examples in an imbalanced class distribution, the classification performance of underrepresented classes is not reflected by the *OA*. In contrast, the class-specific $F1$ scores, being the harmonic means of precision (i.e., the percentage of the images assigned to a certain class that actually corresponds to that class in the reference) and recall (i.e., the percentage of the samples of a class according to the reference which is also assigned to that class by the CNN) reflect the classifier's ability to predict a certain class. We report the average $F1$ scores (also referred to as macro-averaged $F1$ score) per variable, i.e., the average values of all class-specific $F1$ scores of the classes for that variable.

Results

The quality metrics obtained on the validation and test sets are listed in Table 9. These quality metrics are determined on the basis of the prediction results for records (i.e., not on the raw results for individual images in case of records having multiple images). In this section, some general observations and the conclusions drawn from them will be briefly described, where a more detailed analysis of the results can be found in Sect. 6.

Comparing the $F1$ scores as well as the OAs obtained on the validation and the test set, respectively, shows that the hyperparameter tuning on the validation set did not result in overfitting as the order of magnitude of the quality metrics on the validation and the test set are en par. Furthermore, the average $F1$ scores and the OAs are higher in case of minimizing the focal loss in training. Accordingly, it can be concluded that the classifier is able to better predict the classes of the four tasks by focusing on harder training examples, as is realized in the case of the focal loss. In particular, underrepresented classes benefit more from the use of the focal loss, which is indicated by the larger improvements in terms of the $F1$ scores compared to the improvements in terms of OA. The average $F1$ scores over all variables is 3.7% higher in the evaluation for minimizing the focal loss compared to minimizing the softmax cross-entropy, whereas the improvement in terms of OA amounts to 0.9% on average.

Table 10 Hyperparameter tuning

Hyperparameter	Range	Best
Batch size	4, 8, 32 64	4
Learning rate	1e-6, 1e-5, 3e-5, 5e-5 1e-4	1e-5
Weight decay coefficient	0.0, 0.01, 0.02, 0.04 0.05	0.04
Total epochs	4, 8, 12, 16, 20	16

Hyperparameters, the investigated range, and the value chosen to be the best in 50 random trials according to macro-$F1$ evaluated on the validation set

5.2 Text classification

Experimental setup In the text classification experiment we use the method described in Sect. 4.2, implemented using PyTorch [47] and Transformers [60], and the data described in Sect. 3.4 split into training, validation, and test subsets as described in Sect. 3.4. We used the base XLM-R architecture (125M parameters) with 12-layers, 768-hidden-state and its respective provided weights. The layers in the classification heads are initialized using the normal distribution $\mathcal{N}(0.0, 0.02)$ with bias parameters set to zero.

First, we performed a 50-trial random search hyperparameter tuning implemented using Optuna [1]. During hyperparameter tuning, the text classifier is trained on the train set, and we chose the hyperparameters that resulted in the highest macro $F1$ score obtained by evaluating on the validation set. These hyperparameters are detailed in Table 10. We then train a model on the train set with the previously selected hyperparameters and evaluate it on both the validation and test sets.

Table 11 *F*1 scores (*F*1) and overall accuracies (OA) obtained in the multitask experiment both for the validation set and the test set (text classification)

Variable	Validation set		Test set	
	*F*1 [%]	OA [%]	*F*1 [%]	OA [%]
Place	92.2	91.3	93.1	92.6
Timespan	83.8	89.8	82.1	88.0
Technique	87.1	88.2	88.0	89.7
Material	78.5	85.0	78.7	85.4
Average	85.4	88.6	85.5	88.9

Results The results of text classification are shown in Table 11, which presents the overall accuracy and the average *F*1 scores achieved on all records containing text in the validation and test sets. In terms of *F*1 and overall accuracies, these are the best results for any single modality by a significant margin. This is offset by the fact that the text modality is only present in 39.3% of all records (Table 3).

5.3 Tabular classification

Experimental setup The experiments for tabular classification follow a similar protocol as those for the image and text classifiers, the main exception being that this classifier is not based on multitask learning. Thus, for each task we train an individual classifier with different parameters and hyperparameters, selected by task-specific hyperparameter tuning using grid search. We show the hyperparameters, the search space for tuning, and the selected values in Table 12. Note that the ranges selected were all within very reasonable intervals as an additional guard against overfitting.

Results We show the evaluation results in Table 13. Given that it essentially relies on co-occurrences of very coarse labels, the results seem reasonable. In fact, in terms of *F*1 and accuracy, they almost match the image classifier. We also show feature importance by gain in Table 14. For every task, the tabular classifier's most important feature is the museum. That could probably be expected, because museums are not random collections of objects.

Table 13 *F*1 (*F*1) and overall accuracies (OA) obtained in the experiment both for the validation set and the test set (tabular classification)

Variable	Validation set		Test set	
	*F*1 [%]	OA [%]	*F*1 [%]	OA [%]
Place	47.9	62.4	46.2	61.9
Timespan	57.4	65.1	58.6	67.6
Technique	68.6	74.2	68.3	73.0
Material	50.7	82.1	49.4	82.1
Average	55.4	70.0	55.6	71.2

Table 14 Tabular classifier: feature importance per task (information gain)

Target	Feature				
Variable	Museum	Place	Timespan	Technique	Material
Place	0.49	–	0.20	0.12	0.19
Timespan	0.41	0.31	–	0.16	0.12
Technique	0.40	0.29	0.17	–	0.14
Material	0.39	0.21	0.16	0.24	–

5.4 Multimodal classification

Experimental setup For the experiments involving multimodal classifiers, we started by training the three classifiers based on single modalities (images, text, tabular, respectively) on the training set independently of each other in the way described in Sects. 5.1 – 5.3. After that, these classifiers were used to classify the samples in the validation set. Finally, we used these predictions as inputs to train the multimodal classifier on the validation set. We used five-fold cross-validation on the validation set to perform hyperparameter tuning using grid search for the same hyperparameters that used in tuning the tabular classifier. The details of hyperparameter tuning are shown in Table 15.

Results The results of the experiments are shown in Table 16. In this table, we compare the results of the multimodal classifier with and without using the raw tabular data as an additional input. As expected, the variant of

Table 12 Hyperparameter tuning for the tabular classifier: hyperparameters, the investigated range of values (Range) and interval of the search, and best values for each task, chosen by grid search according to macro-*F*1 evaluated on the validation set

Hyperparameter	Range	Interval	Place	Timespan	Technique	Material
colsample_bytree	[0.6, 1.0]	0.2	0.8	0.8	0.6	0.8
Gamma	[0.0, 0.4]	0.2	0.4	0.2	0.2	0.0
learning_rate	[0.1, 0.3]	0.1	0.3	0.3	0.3	0.2
max_depth	[2, 8]	2	4	4	4	8
min_child_weight	[1, 4]	1	1	2	4	2
n_round	[100, 500]	100	100	100	100	500
subsample	[0.6, 1.0]	0.2	0.6	1.0	0.6	0.8

Table 15 Hyperparameter tuning for the multimodal classifier: hyperparameters, the investigated range of values (Range) and interval of the search, and best values chosen by grid search according to macro-$F1$ evaluated on the validation set

Hyperparameter	Range	Interval	Place	Timespan	Technique	Material
colsample_bytree	[0.6, 1.0]	0.2	0.6	0.6	0.8	0.6
gamma	[0.0, 0.4]	0.2	0.2	0.4	0.4	0.4
learning_rate	[0.1, 0.3]	0.1	0.1	0.1	0.3	0.3
max_depth	[2, 8]	2	4	4	6	4
min_child_weight	[1, 4]	1	1	1	4	4
n_round	[100, 200]	100	100	100	100	100
subsample	[0.6, 1.0]	0.2	0.6	0.6	0.6	0.6

The hyperparameter space is the same for all experiments reported in this section. The selected hyperparameters apply to the multimodal classifier using the complete set of input modalities, as shown in Fig. 5 and correspond to the results shown in Table 16

Table 16 $F1$ scores ($F1$) and overall accuracies (OA) on the test set of the multimodal classifier

Variable	$F1$ [%]	OA [%]
Place	76.7	79.0
Timespan	73.1	80.1
Technique	83.8	85.4
Material	61.3	85.5
Average	73.7	82.5

Table 17 Feature importance, measured by information gain, for the multimodal classifier per modality for all tasks

Variable	Text	Image	Tabular
Place	0.47	0.21	0.31
Timespan	0.40	0.23	0.37
Technique	0.20	0.36	0.43
Material	0.46	0.23	0.31

the classifier using these additional input features produces slightly better results than the one without these features. Table 17 gives the feature importance, measured by information gain, of each individual modalitiy. We also performed an ablation study to assess the importance of the individual modalities for the classification results The ablation study was performed by removing one of the modalities from the input of the fusion classifier, leaving only the other two modalities (Table 19).

The overall accuracy and mean $F1$ score achieved by the multimodal classifier are better than those for the image classifier (Table 9) and slightly worse than those reported for the text classifier (Table 11), but this comparison is inconclusive because the results in Table 11 only consider records for which text is available, which is only about 39% of the test set, whereas the evaluation of the image classifier is based on about 96% of the test set and multimodal classification is based on the complete test set.

For the majority of the samples, only images and / or tabular information are available, and thus the prediction would be based on these modalities. To be able to allow for a comparison of the results of all modalities, we carried out an evaluation of all modality-specific classifier and the multimodal classifier on the entire test set. In this evaluation, a record for which a modality was missing was considered a wrong prediction for that modality-specific classifier. For instance, a record without images

Table 19 Average $F1$ scores of the multimodal classifier using input modalities (average over all tasks)

Input modality	$F1$ score [%]
Image + text	70.5
Image + tabular	59.8
Text + tabular	69.9
All	73.7

Table 18 Mean $F1$ scores ($F1$) and overall accuracies (OA) of the different classifiers evaluated on the entire test set

Classifier	Image		Text		Tabular		Multimodal	
Variable	$F1$	OA	$F1$	OA	$F1$	OA	$F1$	OA
Place	38.0	46.8	64.6	42.3	46.2	61.9	76.7	79.0
Timespan	49.2	45.6	54.0	44.7	58.6	67.6	73.1	80.1
Technique	73.5	70.5	40.9	26.1	68.3	73.0	83.8	85.4
Material	46.5	67.5	37.4	21.6	49.4	82.1	61.3	85.5
Average	51.8	57.5	49.2	33.7	55.6	71.2	73.7	82.5

Samples for which a modality was missing are considered as errors for the corresponding modality-specific classifier

was considered to be an incorrect prediction for the image classifier. The resultant overall accuracy values and mean $F1$ scores are shown in Table 18. In this comparison, the multimodal classifier outperforms all the classifiers based on a single modality only.

The results of the modality ablation study, shown in Table 19, and comparing with the individual modality results show in Table 18 confirms the assumption that each modality provides meaningful contribution. The results of a multimodal classifier that combines any two modalities are superior to those of any individual modality. Further, combining all three produces the best results.

6 Analysis

6.1 Image classification

Here, we will provide a detailed analysis of the results of the CNN-based image classifier in Table 9. The table shows that the classification performance strongly varies between tasks. Comparing the OAs, one can see that the variable *material* achieves the highest OAs, followed by *technique* and *timespan*; the worst OA is achieved for the variable *place*. Taking the class structure shown in Table 1 into account, a connection can be made to the number of classes constituting a task's class structure. The larger the number of classes to be distinguished, the lower the achieved percentage of correctly classified images in the softmax experiment, where a similar behavior can be observed for the focal experiment; *material* having three classes has the highest OA of 80.7%, followed by *technique* having four classes with 76.8% correct predictions and *timespan* with five classes with a OA of 64.0%, whereas *place* with nine classes has the lowest OA of 62.2%.

An analysis of the task-specific $F1$ scores in connection with the class distributions of the respective task indicates a dependency of the $F1$ score on the degree of class imbalance. Taking the ratio of the number of image examples for the majority class, i.e., the class with the most labeled examples in the dataset, in relation to the number of image examples for the minority class, i.e., the class with the fewest examples, a negative correlation between this ratio and the achieved task-specific $F1$ score can be observed for the focal loss experiment, where a similar behavior can be observed for the softmax experiment. The majority class of *technique* has 2.5 times as many examples as the minority class and *technique* has the highest $F1$ score of 77.9%, followed by *timespan* with a ratio of 4.9 and a score of 57.5% and *material* with a ratio of 7.7 having a score of 51.2%. The lowest $F1$ score of 47.0% is obtained for *place* with a ratio of 8.3. We attempted to overcome this dependency of the $F1$ scores on the class distributions through focusing on hard training examples by means of the presented variant of the focal loss in Eq. 2. Analyzing the improvements of the $F1$ scores by utilizing the focal loss instead of the softmax cross-entropy loss shows that the focal loss indeed reduces this dependency: except for the variable *place*, there is an improvement of the task-specific $F1$ scores, and in these cases it is larger for tasks with a high class imbalance (indicated by a high ratio between the number of examples for the majority class and the minority class, respectively). The $F1$ score of *material* (ratio of 7.7) is improved by 7.8%, whereas the $F1$ score of *technique* (ratio of 2.5) is improved by 3.9%. The variable *place* with a ratio of 8.3 should have received the largest improvement in $F1$ score according to the general trend, but it actually is slightly worse (-0.2%). We assume this to be related to the large number of classes to be distinguished for *place*, which might make a correct prediction more complicated for this variable than for the other ones.

In summary, the utilization of the focal loss improves the performance of the trained classifier in correctly predicting the properties of silk based on images. Even though the variable-specific $F1$ score still seems to depend on the degree of imbalance of a task's class distribution, focusing on hard examples during training primarily improves the task-specific $F1$ scores of tasks with large class imbalances, as long as the number of classes to be differentiated is not too large. Solving the remaining challenge of predicting all classes of a task equally well may require more data, as not all aspects of all silk properties are equally well represented in the available images.

6.2 Text classification

The results for the text modality, shown in Table 11, are better, in terms of $F1$ and OA, than the results for the image modality (Table 9) or the tabular modality (Table 13). This is not surprising, since the properties we are predicting are important to domain experts, they, or their taxonomy subclasses, are often included in the text descriptions of the cultural heritage objects (altough not necessarily using the same words). Even when they are not, we can intuitively expect some degree of similarity between text descriptions of objects with similar underlying values for these properties, either globally or at least within the same museum.

The biggest disadvantage of the text classifier is that text descriptions are present in the dataset far less often than images, as shown in Table 3. Once adjusted for missing modality, given that more than 60% of the records are missing this modality, the text classifier actually performs the worst of any modality, as shown in Table 18.

We analyzed about 20 misclassified English language test set examples for each task. In around half of the cases, there was no direct information that could've allowed an accurate classification. E.g., no location mentioned when attempting to classify place or no year mentioned when attempting

Table 20 Examples of misleading text descriptions

Text description (Snippet)	Predicted	True
Motifs found on **seventeenth-century** coverlets but must have been made in the early eighteenth century (...) Embroidery of Gujerat [sic] in the **Seventeenth Century**	XVII	XVIII
Derived from engravings after Maarten de Vos which first appeared in Gerard de Jode's **1579** illustrated bible	XVI	XVII
Center text reads **"Vole vole mon coeur!**	FR	GB
Depiction from the **Italian** poem	IT	GB

Emphasis added to highlight the misleading text snippets

to classify timespan. This forces the classifier to rely on other statistical regularities present in the text to provide a classification.

The *material* task is particular. Its most common class, "animal fibre", is a de facto background class. All records in the dataset should be of silk fabrics, which means the material they are made of is an "animal fibre". Some have other materials too. These other materials can correspond to a vegetable fibre (e.g., cotton) or a thread with some metal (e.g., gold thread). While the problem of not having label specific information in the text is common in the examples we analyzed (6/20), obviously incorrectly labeled examples were even more common (9/20). This occurs when either the original record was missing the correct label or when the automatic extraction and linking of the label failed. The high prevalence of this type of error within this task in the examples we analyzed, combined with its absence in other tasks, leads us to suspect that this is the main cause of the relatively lower accuracy and $F1$ scores for this task.

The *technique* task is also particular in terms of the examples we analyzed. A significant number of examples (5/19) contain information that would imply multiple labels, where usually a small part of the object was produced using a different technique from the main part of the object. A similar type of error occurs in the *timespan* task within a similar proportion of examples (5/10). In the *timespan* task, this can occur when an object was produced at a certain date but later altered or when the estimated date of production within the text crossed centuries.

We hypothesize that the somewhat better results for the place task are connected to regularities between the museum and an object's place of production. This connection is suggested in Table 14. Text descriptions are very indicative of the museum, not just in the language but usually also in style, length, and topics.

Finally, we would like to point out at the existing of misleading text examples such as in Table 20 where information in the text can correspond to an incorret label. We do this to give the reader a better understanding of the challenges faced by an algorithm.

6.3 Tabular classification

When compare to the other modalities, the tabular classifier performs (Table 13) slightly worse than the image classifier (Table 9) in terms of average $F1$ score, respectively, 55.6% versus 58.4%. However, this situation is reversed when accounting for missing modality (Table 18), with the tabular classifier outperforming the image classifier with $F1$ scores of 55.6% versus 51.8%. Intuitively, from a domain perspective, we can expect that these variables to be associated. For example, a certain country is more active in the textile industries during a certain timespan than during others. Further, museums are typically curated and not random collections. However, given the limited number of features and the coarseness of the labels, we should not overestimate the strength of the association between variables, which we calculated as Cramer's V in Fig. 6. Cramers' V is a symmetric measure that gives a value between 0 and 1 for the association between two nominal variables. We see that the values are, by themselves, relatively low, with

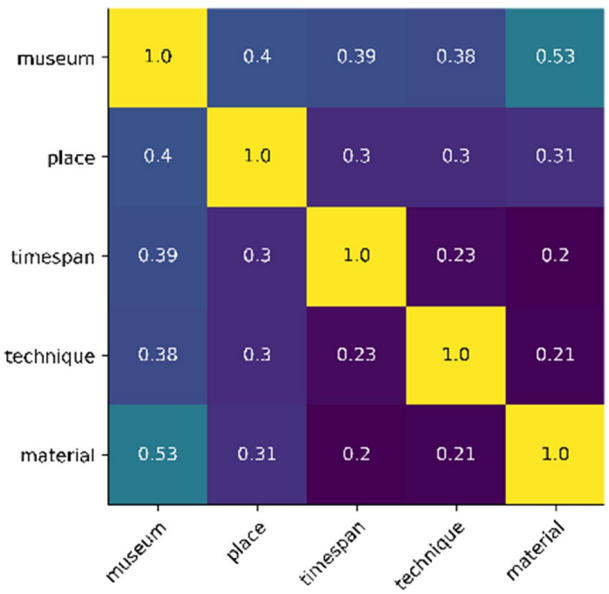


Fig. 6 Association between features of the tabular classifier (Cramer's V)

only material and museum having a value above 0.5. The association between museum and the other properties was the highest (first column or row) which again reinforces the belief that the curated nature of museum collections and its impact on the other properties is learnable from this dataset. The strength of the association does not directly translate into feature importance as measured by information gain (Table 14). Rather, it seems, a particular combination of these associations is being learned by the tree boosting algorithm in such a way that results in a relatively effective classifier.

6.4 Multimodal classification

As pointed out in Sect. 5.4, the comparison of the classification accuracies indicates that the text classifier achieves the best performance of all modalities (Table 11), but of course it is only applicable when text is available, which is only the case for a relatively low number of records, (cf. Table 3). This confirms our hypothesis that multimodal classification results in a better classification performance if one of the aspects under consideration is to obtain correct predictions for a number of records that is as large as possible. When evaluated on samples having text, the text classifier might achieve higher accuracy metrics; however, a considerable percentage of samples cannot be classified in that way, and the total number of correct classifications is largest when using multimodal classification (cf. Table 18).

As Tables 17, 18 and 19 imply, each modality contributes significantly to the multimodal classifier. Looking at the feature importance of the multimodal classifier in Table 17, we can see that the output of the text classifier is the most important feature, except when it comes to predicting technique, almost certainly due to the relatively small number of records with an annotation for *technique* in the validation set for which text is available (487 records as opposed to about 1100-1600 for the other tasks). Most errors in the timespan task occur between chronologically similar dates. Most errors in the place task occur between countries that are geographically close to each other, e.g., Italy (IT) and France (FR). As far as material is concerned, errors occur primarily between animal fiber and the other labels, because all objects are made of silk and due to the label imbalance. No clear trend can be observed for the prediction of technique.

7 Integrating and visualizing the predictions

To model the predictions as part of the SILKNOW Knowledge Graph ontology we use classes and properties of the Provenance Data Model (Prov-DM), more specifically the PROV ontology (PROV-O)[10], an OWL2 ontology. It makes it possible to map PROV-DM to RDF. It allows the expression of important elements of the predictions for each modality. These different predictions can be represented using different `prov:activity` classes each. The image, text description, or categories each prediction is based on are represented with the property `prov:used`. The exact date of the prediction is represented with prov:atTime and prov:wasAssociatedWith connects the activity class to the `prov:SoftwareAgent` class, which is used to describe the particular algorithm and model used. The predicted metadata value is represented with rdf:Statement, connected to prov:activity via a `prov:wasGeneratedBy` property. The confidence score of the prediction is expressed through the property L18 ("has confidence score") from our own SILKNOW Ontology. The predicted value is expressed in form of a URI with `rdf:object`, the type of the predicted property through rdf:predicate and its fitting CIDOC-CRM property type. The property `rdf:subject` connects the statement to the production class (E12) of the object in the Knowledge Graph. Every prediction is inserted in the appropriate part of the existing KG. For example, if a material value gets predicted, it gets inserted with the CIDOC-CRM property `P126_employed` at the production class of the object. See Fig. 7 for an illustration of the data model.

The prediction models were only trained on group labels (Sect. 3.3), and thus can only predict those. It is sometimes necessary to map them back to a more concrete concept of the SILKNOW Thesaurus. If, for example, "Damask" is predicted in form of its facet link http://data.silknow.org/vocabulary/facet/damask it will be automatically converted into http://data.silknow.org/vocabulary/168, as facet links are too general for concrete category values. All predictions are converted one after another using the described data model and saved as a turtle file format which is uploaded and stored as its own graph identified by http://data.silknow.org/predictions. This makes it possible to always identify and eventually separate predictions from original values obtained from the museums. In total, 98,379 predictions were made for 19,248 distinct objects. These were uploaded into the SILKNOW Knowledge Graph.

In our exploratory search engine ADASilk, the predictions are displayed differently from values that come originally from the museums: They are shown in blue together with their confidence score as a percentage next to them. A tooltip is available to explain how this value was predicted, including the modality, algorithm, model identifier, etc. To display predictions like this on ADASilk, the respective SPARQL query was updated and new subqueries are used to take into account the aforementioned new properties. See Fig. 8 for a screenshot.

[10] https://www.w3.org/TR/prov-dm/.

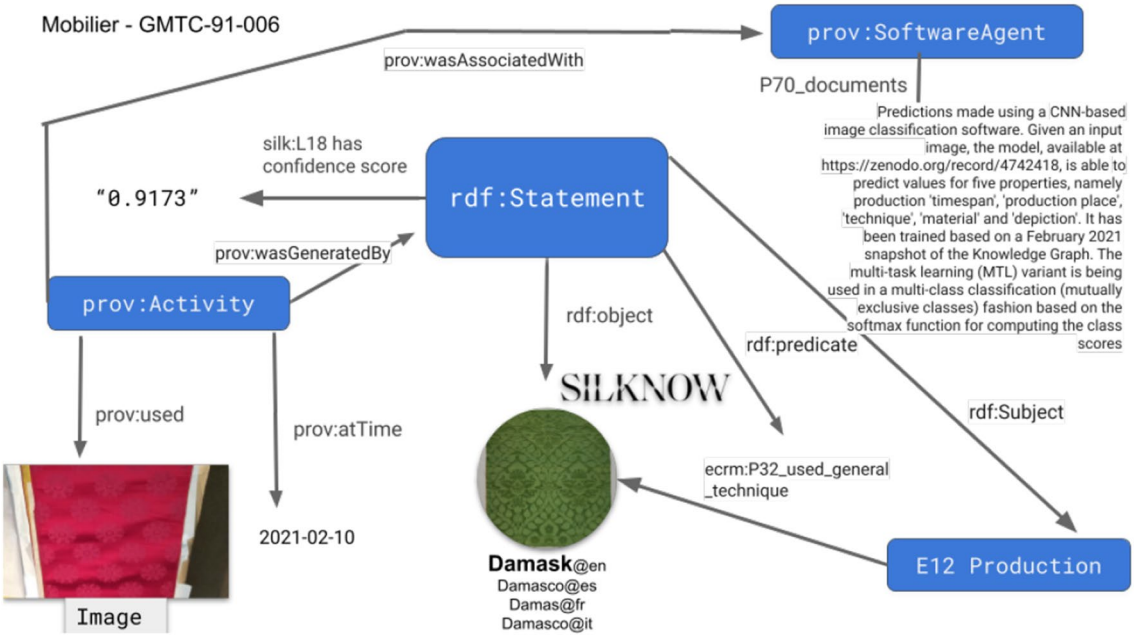


Fig. 7 Graph showing the prediction of the production technique (damask) with a high confidence score (0.9173) using the textual analysis software

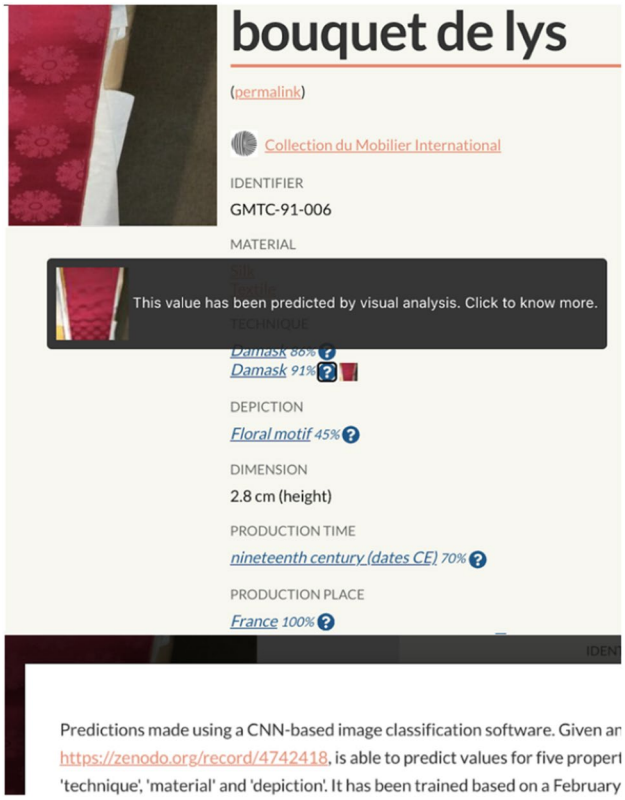


Fig. 8 UI Screenshot showing the prediction of the production technique (damask) with a high confidence score (0.9173) using the image analysis software

8 Conclusions

We presented results for three individual modality classifiers, as well as multimodal results. In terms of our original hypothesis, presented in Sect. 1.2, we showed that we were in fact able to accurately predict missing properties in the digitized silk fabric artifacts that made up our dataset. While the quality of predictions varied between individual modalities, we showed that the multimodal approach provided the best results. To recapitulate, our contributions included the already mentioned multimodal approach tailored specifically to the challenges we faced in the multimodal scenario, including the incomplete overlap of data across modalities. The individual approaches of each modality-specific classifier also provide a useful contribution to the automated classification of cultural heritage objects. The image and text classifier offer the possibility of being applied to data outside a Knowledge Graph (KG) or database, possibly even directly submitted by the user of a system. The tabular classifier, on the other hand, offers the possibility of classifying data in a KG or database when no text descriptions or images are present by relying on other properties. It is also important to remember that in most practical situations, including inside a KG or other knowledge bases, images are more common than text descriptions of objects in the cultural heritage domain.

The data we used in our work originally comes from many museum sources and is from a very specific Cultural Heritage domain: historical, European silk fabrics. We

applied common methods to process such data and developed an ontology and a Knowledge Graph out of the original museum texts and images. Such an effort comes typically with challenges, which in our case consisted mostly of a small amount of (training) data, domain specificity, different styles of writing texts and capturing images of objects, different languages (in the case of texts) and finally simply annotation errors, typos, and other errors that happened during the original digitization. Not all of these challenges can be completely overcome and some of them, like the metadata gaps, even constitute part of the motivation to conduct this research work. As some data imperfection could still not be totally excluded, some removal of data was necessary to ensure sufficiently clean and class balanced data for our supervised approaches. This could, however, be very much alleviated through the grouping of certain labels, which was also possible through our domain expert-designed thesaurus about silk fabric concepts. In the end we can present a cultural heritage dataset that can be used for automated classification or even multimodal approaches. In this work, we also provide the data modelling of how metadata predictions for data such as our can be represented within knowledge graphs or other knowledge bases.

We have shown that properties of silk fabrics can be predicted from images of these fabrics. In this context, we proposed to use the focal loss for training to compensate for the effects of class imbalance in the training set, a problem that is quite common in the cultural heritage domain. Our results indicate that the proposed strategy can mitigate this problem to a certain degree, in particular improving the classification performance for the underrepresented classes in terms of the $F1$ score. Image classification performs particularly well for the task to predict the technique used for producing a fabric. Nevertheless, there is still room for improvement, as indicated by the performance metrics for all variables.

When text descriptions are present, the text classifier provides the best results of any single modality. It seems, thus, that the text classifier was able to overcome the primary challenges it faced: small dataset, domain specificity, cross-linguality, and museum specific text styles. This was primarily achieved by the choice of XLM-R as the basis of the text classifier. Misleading text descriptions stand out as a challenge for text classification.

When all data is considered, we have shown that the multimodal approach is the best according to the macro $F1$ metric. While most records contain images, not all do (3.4%) and a smaller number of records contain neither text nor images (2.1%). On the other hand, if we had tried to implement a classifier using the text modality alone, we could only classify 40% of the records. While we can say that a multimodal approach does allow us to classify a greater number of records than using images alone, the primary practical benefit of the multimodal approach over performing just image classification is probably the qualitative improvement in classification results demonstrated.

In terms of the dataset, a perhaps a better approach could be found for dealing with noisy labels, as well as finding better ways to deal with fine-grained labels and label ontology mismatches.

Future work on image classification could concentrate on improving the performance for underrepresented classes even more, e.g., by using methods for few-shot learning. Furthermore, as some experimental results indicated that some training labels might be incorrect, training methods that are robust against such errors ("label noise") could be investigated.

The code and data used to perform the experiments reported in this work is available online at https://github.com/silknow/multimodal_cultural_heritage and https://zenodo.org/record/6590957, respectively.

Acknowledgements This work was supported by the Slovenian Research Agency and the European Union's Horizon 2020 research and innovation program under SILKNOW grant agreement No. 769504.

References

1. Akiba, T., Sano, S., Yanase, T., et al.: Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD International conference on knowledge discovery and data mining (2019)
2. Arik, S.O., Pfister, T.: Tabnet: Attentive interpretable tabular learning. In: Proceedings of the AAAI conference on artificial intelligence 35(8):6679–6687. (2021) https://ojs.aaai.org/index.php/AAAI/article/view/16826
3. Arora, R.S., Elgammal, A.M.: Towards automated classification of fine-art painting style: a comparative study. In: International conference on pattern recognition, pp. 3541–3544 (2012)
4. Belhi, A., Bouras, A., Foufou, S.: Leveraging known data for missing label prediction in cultural heritage context. Appl. Sci. (2018). https://doi.org/10.3390/app8101768
5. Belhi, A., Bouras, A., Foufou, S.: Towards a hierarchical multitask classification framework for cultural heritage. In: 2018 IEEE/ACS 15th international conference on computer systems and applications (AICCSA), IEEE, pp. 1–7 (2018b)
6. Bishop, C.M.: Pattern Recognition and Machine Learning, 1st edn. Springer, New York (NY), USA (2006)
7. Blessing, A., Wen, K.: Using machine learning for identification of art paintings. Technical report. Stanford University, USA (2010)
8. Bojanowski, P., Grave, E., Joulin, A., et al.: Enriching word vectors with subword information. Trans. Assoc. Comput. Linguist. **5**, 135–146 (2017)
9. Caruana, R.: Multitask learning. Mach. Learn. **28**(1), 41–75 (1997)
10. Caruana, R.A.: Multitask learning: a knowledge-based source of inductive bias. In: International conference on machine learning, pp. 41–48 (1993)
11. Castellano, G., Vessio, G.: Deep learning approaches to pattern extraction and recognition in paintings and drawings: an overview. Neural Comput. Appl. **33**(19), 12263–12282 (2021)

12. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. Association for computing machinery, New York, NY, USA, pp. 785–794, https://doi.org/10.1145/2939672.2939785(2016)
13. Conde, M.V., Turgutlu, K.: Clip-art: contrastive pre-training for fine-grained art classification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR) workshops, pp. 3956–3960 (2021)
14. Conneau, A., Khandelwal, K., Goyal, N., et al.: Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th annual meeting of the association for computational linguistics. Association for computational linguistics, Online, pp. 8440–8451, https://doi.org/10.18653/v1/2020.acl-main.747 (2020)
15. Crawshaw, M.: Multi-task learning with deep neural networks: a survey. arXiv preprint arXiv:2009.09796 (2020)
16. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), IEEE, pp 886–893 (2005)
17. Deng, J., Dong, W., Socher, R., et al.: Imagenet: a large-scale hierarchical image database. In: IEEE conference on computer vision and pattern recognition, pp. 248–255 (2009)
18. Devlin, J., Chang, M.W., Lee, K., et al.: 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 (Long and Short Papers). Association for computational linguistics, Minneapolis, Minnesota, pp. 4171–4186, https://doi.org/10.18653/v1/N19-1423 (2019)
19. Doerr, M.: The CIDOC CRM, an ontological approach to schema heterogeneity. In: Semantic interoperability and integration (2005)
20. Dorozynski, M., Clermont, D., Rottensteiner, F.: Multi-task deep learning with incomplete training samples for the image-based prediction of variables describing silk fabrics. ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. **4**(2/W6), 47–54 (2019)
21. Fiorucci, M., Khoroshiltseva, M., Pontil, M., et al.: Machine learning for cultural heritage: a survey. Pattern Recogn. Lett. **133**, 102–108 (2020)
22. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Ann. Stat. **29**(5), 1189–1232 (2001)
23. Gao, Y., Li, Y., Lin, Y., et al.: Deep learning on knowledge graph for recommender system: a survey. CoRR abs/2004.00387. arXiv:2004.00387 (2020)
24. Garcia, N., Vogiatzis, G.: How to read paintings: semantic art understanding with multi-modal retrieval. In: Proceedings of the European conference in computer vision workshops (2018)
25. Garcia, N., Renoust, B., Nakashima, Y.: Contextnet: representation and exploration for painting classification and retrieval in context. Int. J. Multimed. Inf. Ret. **9**(1), 17–30 (2020)
26. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778 (2016a)
27. He, K., Zhang, X., Ren, S., et al.: Identity mappings in deep residual networks. In: European conference on computer vision, pp. 630–645 (2016b)
28. Hyvönen, E., Mäkelä, E., Kauppinen, T., et al.: Culturesampo: a national publication system of cultural heritage on the semantic web 2.0. In: Aroyo, L., Traverso, P., Ciravegna, F., et al. (eds.) The Semantic Web: Research and Applications, pp. 851–856. Springer, Berlin Heidelberg (2009)
29. Iqbal Hussain, M.A., Khan, B., Wang, Z., et al.: Woven fabric pattern recognition and classification based on deep convolutional neural networks. Electronics **9**(6), 1048 (2020)
30. Joulin, A., Bojanowski, P., Mikolov, T., et al.: Loss in translation: learning bilingual word mapping with a retrieval criterion. In: Proceedings of the 2018 conference on empirical methods in natural language processing (2018)
31. Kadra, A., Lindauer, M., Hutter, F., et al.: Well-tuned simple nets excel on tabular datasets. In: Beygelzimer A., Dauphin Y., Liang P., et al. (eds) Advances in Neural Information Processing Systems, https://openreview.net/forum?id=d3k38LTDCyO (2021)
32. Kingma, DP., Ba, J. Adam: A method for stochastic optimization. In: 3rd International conference on learning representations (ICLR 2015) (2015a)
33. Kingma, DP., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (Poster), arXiv:1412.6980 (2015b)
34. Koch, I., Ribeiro, C., Lopes, C.: ArchOnto, a CIDOC-CRM-based linked data model for the Portuguese archives, Springer, pp 133–146. https://doi.org/10.1007/978-3-030-54956-5_10(2020)
35. Krizhevsky, A., Sutskever, I., Hinton, GE.: ImageNet classification with deep convolutional neural networks. In: Advances in neural information processing systems 25 (NIPS'12), pp 1097–1105 (2012)
36. LeCun, Y., Boser, B., Denker, J.S., et al.: Backpropagation applied to handwritten ZIP code recognition. Neural Comput. **1**(4), 541–551 (1989)
37. Li, X., Chen, C.H., Zheng, P., et al.: A knowledge graph-aided concept-knowledge approach for evolutionary smart product-service system development. J. Mech. Des. **142**(101), 403 (2020). https://doi.org/10.1115/1.4046807
38. Lin, T.Y., Goyal, P., Girshick, R., et al.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision, pp. 2980–2988 (2017)
39. Liu, W., Chen, L., Chen, Y.: Age classification using convolutional neural networks with the multi-class focal loss. IOP Conf. Ser.: Mater. Sci. Eng. **428**(012), 043 (2018). https://doi.org/10.1088/1757-899x/428/1/012043
40. Liu, X., He, P., Chen, W., et al.: Multi-task deep neural networks for natural language understanding. In: Proceedings of the 57th annual meeting of the association for computational linguistics. Association for computational linguistics, Florence, Italy, pp. 4487–4496. https://doi.org/10.18653/v1/P19-1441 (2019a)
41. Liu, Y., Ott, M., Goyal, N., et al.: Roberta: a robustly optimized Bert pretraining approach. 1907.11692 (2019b)
42. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: 7th International conference on learning representations, ICLR 2019, New Orleans, LA, USA, May 6–9 (2019)
43. Meng, S., Pan, R., Gao, W., et al.: A multi-task and multi-scale convolutional neural network for automatic recognition of woven fabric pattern. J. Intell. Manuf. **32**(4), 1147–1161 (2021)
44. Mensink, T., Van Gemert, J.: The Rijksmuseum challenge: museum-centered visual recognition. In: Proceedings of international conference on multimedia retrieval, pp. 451–454 (2014)
45. Nair, V., Hinton, G.E.: Rectified linear units improve restricted Boltzmann machines. In: Proceedings of the 27th international conference on machine learning (ICML-10), pp 807–814 (2010)
46. Palagi, E.: Evaluating exploratory search engines: designing a set of user-centered methods based on a modeling of the exploratory search process. PhD thesis, Université Côte d'Azur (2018)
47. Paszke, A., Gross, S., Massa, F., et al.: Pytorch: An imperative style, high-performance deep learning library. In: Wallach H., Larochelle H., Beygelzimer A., et al (eds) Advances in Neural Information Processing Systems 32. Curran Associates, Inc., pp. 8024–8035 (2019). http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf
48. Puarungroj, W., Boonsirisumpun, N.: Recognizing hand-woven fabric pattern designs based on deep learning. In: Advances in Computer Communication and Computational Sciences, pp. 325–336. Springer, Singapore (2019)

49. Radford, A., Kim, JW., Hallacy, C., et al.: Learning transferable visual models from natural language supervision. In: Meila M., Zhang T. (eds) Proceedings of the 38th international conference on machine learning, proceedings of machine learning research, vol. 139. PMLR, pp. 8748–8763 (2021)
50. Ruder, S.: An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098 (2017)
51. Ruotsalo, T., Aroyo, L., Schreiber, G., et al.: Knowledge-based linguistic annotation of digital cultural heritage collections. IEEE Intell. Syst. **24**(2), 64 (2009)
52. Santos, I., Castro, L., Rodriguez-Fernandez, N., et al.: Artificial neural networks and deep learning in the visual arts: a review. Neural Comput. Appl. **33**(1), 121–157 (2021)
53. Sharif Razavian, A., Azizpour, H., Sullivan, J., et al.: CNN features off-the-shelf: an astounding baseline for recognition. In: IEEE Conference on computer vision and pattern recognition workshops, pp 806–813 (2014)
54. Srivastava, N., Hinton, G., Krizhevsky, A., et al.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(2014), 1929–1958 (2014)
55. Stefanini, M., Cornia, M., Baraldi, L., et al.: Artpedia: a new visual-semantic dataset with visual and contextual sentences. In: Proceedings of the international conference on image analysis and processing (2019)
56. Strezoski, G., Worring, M.: Omniart: multi-task deep learning for artistic data analysis. arXiv preprint arXiv:1708.00684 (2017)
57. Sur, D., Blaine, E.: Cross-depiction transfer learning for art classification. Tech. Rep. CS 231A and CS 231N, Stanford University, USA (2017)
58. Tan, WR., Chan, C.S., Aguirre, HE., et al.: Ceci n'est pas une pipe: a deep convolutional network for fine-art paintings classification. In: IEEE International conference on image processing, pp. 3703–3707 (2016)
59. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. In: Guyon I., Luxburg U.V., Bengio S., et al (eds) Advances in Neural Information Processing Systems, vol 30. Curran Associates, Inc., https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (2017)
60. Wolf, T., Debut, L., Sanh, V., et al.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations. Association for computational linguistics, Online, pp 38–45, https://www.aclweb.org/anthology/2020.emnlp-demos.6 (2020)
61. Xiao, Z., Liu, X., Wu, J., et al.: Knitted fabric structure recognition based on deep learning. J. Text. Inst. **109**(9), 1–7 (2018)
62. Yosinski, J., Clune, J., Bengio, Y., et al.: How transferable are features in deep neural networks? Adv. Neural Inf. Process. Syst. **27**, 3320–3328 (2014)
63. Zhang, C., Kaeser-Chen, C., Vesom, G., et al.: The imet collection 2019 challenge dataset. arXiv preprint arXiv:1906.00901 (2019)
64. Zou, X.: A survey on application of knowledge graph. J. Phys: Conf. Ser. **1487**(012), 016 (2020). https://doi.org/10.1088/1742-6596/1487/1/012016

Chapter 4

Semi-Supervised Learning

This chapter includes the work published in the article titled *Detecting Fine-Grained Emotions in Literature* (Rei & Mladenić, 2023). The article was published in the Special Issue Natural Language Processing (NLP) and Applications of the Applied Sciences journal, Volume 13, in July 2023. The journal has an impact factor of 2.7 and is indexed in the Science Citation Index Expanded. In our previous work (Pita Costa et al., 2021; Rei et al., 2023; Rei & Mladenić, 2020), we leveraged transfer learning and thesaurus mapping to create our datasets. In this work, we define our own set of 38 emotion labels and use semi-supervised learning (SSL) methods to create a multi-label emotion detection from text dataset, using sentences from literature. Our methodology uses textual entailment zero-shot classification to perform emotion-specific weak-labeling, selecting examples with the highest and lowest scores from a large corpus: the English books in Project Gutenberg. Utilizing these emotion-specific datasets, we train binary classifiers for each individual emotion and use these binary classifiers to label all examples, and thus construct a multi-label dataset. The model, which uses RoBERTa (Y. Liu et al., 2019), fine-tuned on this dataset produced good results when evaluated on a manually labeled gold set with a macro averaged F1 score of 59%. This compares favorably to crowdsourced datasets.

Furthermore, we conduct evaluations of the model's performance in transfer scenarios using benchmark datasets including both source-only transfer and fine-tuning few-shot transfer. With source-only transfer we reach an F1 of 74% on the Tales dataset, while an equivalent model trained directly on that dataset reaches an F1 of 83%. This means that with source-only transfer we achieve 93% of the performance of directly training on that dataset. The same experimental setting evaluated on the EMOINT (Mohammad & Bravo-Marquez, 2017) social media dataset leads to a 60% F1 compared to 82% for an equivalent model trained directly on that dataset. Surprisingly, this result is still better than the same source-only transfer from a different social media dataset. In the fine-tuning few-shot transfer experiments, we consistently achieve significantly better results than the no-transfer baseline for every dataset when training with up to 150 target dataset examples. At 150 examples, we achieve an F1 of 79% compared to the baseline 64% on Tales and 70% compared to the baseline 47% on EMOINT.

The results indicate that we were able to build an emotion detection dataset without manual annotation and that the knowledge learned from our dataset exhibits transferability to different domains such as social media. Key differences between our work and previous multi-label emotion datasets include the fact that we used a semi-supervised approach instead of crowdsourcing, we have considerably more labels than previous datasets, and it is the first multi-label emotion dataset for literature. Finally, we also included an analysis of emotion correlation within the dataset which shows that contrasting emotions within a single sentence is common in literature, which is a valuable conclusion for future work.

applied sciences

MDPI

Article

Detecting Fine-Grained Emotions in Literature

Luis Rei [1,2,*] and Dunja Mladenić [1]

[1] Jožef Stefan Institute, 1000 Ljubljana, Slovenia; dunja.mladenic@ijs.si
[2] Jožef Stefan International Postgraduate School, 1000 Ljubljana, Slovenia
* Correspondence: luis.rei@ijs.si

Abstract: Emotion detection in text is a fundamental aspect of affective computing and is closely linked to natural language processing. Its applications span various domains, from interactive chatbots to marketing and customer service. This research specifically focuses on its significance in literature analysis and understanding. To facilitate this, we present a novel approach that involves creating a multi-label fine-grained emotion detection dataset, derived from literary sources. Our methodology employs a simple yet effective semi-supervised technique. We leverage textual entailment classification to perform emotion-specific weak-labeling, selecting examples with the highest and lowest scores from a large corpus. Utilizing these emotion-specific datasets, we train binary pseudo-labeling classifiers for each individual emotion. By applying this process to the selected examples, we construct a multi-label dataset. Using this dataset, we train models and evaluate their performance within a traditional supervised setting. Our model achieves an F1 score of 0.59 on our labeled gold set, showcasing its ability to effectively detect fine-grained emotions. Furthermore, we conduct evaluations of the model's performance in zero- and few-shot transfer scenarios using benchmark datasets. Notably, our results indicate that the knowledge learned from our dataset exhibits transferability across diverse data domains, demonstrating its potential for broader applications beyond emotion detection in literature. Our contribution thus includes a multi-label fine-grained emotion detection dataset built from literature, the semi-supervised approach used to create it, as well as the models trained on it. This work provides a solid foundation for advancing emotion detection techniques and their utilization in various scenarios, especially within the cultural heritage analysis.

Keywords: emotion detection; semi-supervised learning; weak-labeling; pseudo-labeling; benchmark literature dataset

Citation: Rei, L.; Mladenić, D. Detecting Fine-Grained Emotions in Literature. *Appl. Sci.* **2023**, *13*, 7502. https://doi.org/10.3390/app13137502

Academic Editor: Valentino Santucci

Received: 29 May 2023
Revised: 19 June 2023
Accepted: 22 June 2023
Published: 25 June 2023

1. Introduction

Emotions are an integral part of human beings. Although prominent in neuroscience, psychology, and behavioral science, it is studied in the context of other fields, including healthcare, industrial design, architecture, and computer science, particularly in the context of Affective Computing [1]. The communication of emotion forms an integral part of human interaction but is often hard to convey in text. As humans shifted towards communicating over text messages, the need to better convey emotions was obvious. This led to the widespread adoption of emoticons and later emojis. These have even been adopted by chatbot developers to more effectively communicate with their users. However, humans have been communicating through text long before the modern age. Emotion has long been a crucial aspect of storytelling, including writing and understanding literature [2–4]. Even if, the particular set of emotions expressed can change significantly over time [5].

1.1. Motivation

Emotion detection in cultural heritage, literature in particular and not limited to fiction, is the focus of this work. For our immediate motivation and near-future application, this work follows Ref. [6] which combined emotion detection together with the detection of olfactory stimuli to begin associating certain emotions with sensory stimuli. The authors

used only five emotions because their emotion detection model was trained on the Tales dataset [7]. We want to allow this type of analysis of cultural heritage to be conducted with fine-grained emotions.

The most relevant dataset for emotion detection in literature, Tales, consists of 1207 sentences extracted from fairy tales obtained from Project Gutenberg (https://www.gutenberg.org/ (accessed on 22 June 2023)) and originally annotated with one of Ekamn's six basic emotions [8]: anger, fear, disgust, happiness, sadness, and surprise. Later these were reduced to just five emotions by merging anger and sadness. This brings to immediate attention the shortcomings we hope to address:

1. The use of coarse-grained emotions stemming from the use of "basic" or "top level" categories in psychological models of emotions;
2. A single emotion label per sentence;
3. The small size of the dataset by current machine learning standards.

Coarse-grained "basic" emotions are not sufficient for any application or study that requires identifying a more specific emotion. For example, a generic "sadness" emotion category can include distinct emotional states such as grief, despair, disappointment, and nostalgia. Further, it is debatable whether these emotion categories are appropriate at all for annotating text and whether their use, over specific emotions, complicates the annotation task [9,10].

With regard to datasets and models with a single emotion label per sentence, the question becomes what to do if a sentence expresses multiple emotions. For example, happiness and sadness can be expressed in the same sentence: "They celebrated their final reunion, knowing that their paths would soon diverge, leaving behind a profound sense of joy tinged with the ache of farewell.". This has led to more recent work focusing on multi-label emotion detection (e.g., Refs. [11–13]).

Cultural Heritage, such as the study literature, is a knowledge-intensive domain. Datasets that focus on literature typically require expert or near-expert annotators (e.g., graduate students and their advisors) [14,15], or at least, more substantial efforts beyond simple crowdsourcing. This often results in small, imbalanced datasets [16]. We believe the approach we propose might also be used in other problems and domains that so far have relied on small expert annotated datasets or zero-shot transfer, such as healthcare [17].

1.2. Overview

To address the issues mentioned above, we introduce a large multi-label fine-grained emotion detection dataset. To create it, and overcome the difficulty in annotating literature, we propose to use a semi-supervised learning approach. We begin with sentences from Project Gutenberg (Section 2.1 which we preprocess and de-duplicate (Section 2.2. We create a fine-grained emotion taxonomy (Section 2.3 and leverage Natural Language Inference (abbreviated NLI, and also known as Textual Entailment) [18] for weak supervision (Section 2.4) through zero-shot text classification [19], picking the highest scored sentences for each emotion. In the next step of our approach, we train a binary classifier for each emotion based on the weak labels and use each of them to perform pseudo-labeling for all examples. In this way, from binary weakly labeled examples, we create a multi-label dataset through pseudo labeling (Section 2.5). Since we use emotion-specific binary classifiers to create the final labels of the dataset, the probability assigned to each emotion by its respective classifier can be used as a soft label or converted to a hard label with a threshold (probability ≥ 0.5). We provide a detailed analysis of the dataset in Section 3, including the label distribution (Section 3.1), label correlation (Section 3.2), and label sentiment (Section 3.3). Finally, we train two supervised models, described in Section 2.6, on this dataset, with the main purpose of assessing the quality of our dataset. To evaluate our dataset and models, we begin with the typical supervised evaluation scenario against a human-labeled subset of our dataset (Section 4.2). Next, we evaluate models trained on our dataset in a zero-shot transfer learning setting by mapping the output of our classifier (Section 4.3). As our last experiment, we evaluate few-shot transfer learning based on

fine-tuning the models trained on our dataset (Section 4.4). In Section 5, we discuss our results. Finally, in Section 6 we present our conclusions, where we see the biggest room for improvement, and our plans for future work.

Our primary research objective is to explore the effectiveness of a semi-supervised approach using NLI to create a fine-grained multi-label dataset for training supervised emotion detection models specialized for literature. This approach aims to tackle the challenges presented by the complex nature of the problem and the domain, which hinder accurate annotation by crowdworkers and incur high costs for expert annotation. Our secondary research objective is to provide a dataset and associated taxonomy model that includes a set of emotion labels not present in previous work and is geared towards literature analysis.

1.3. Related Work

Annotating training data for machine learning, especially in knowledge-intensive contexts, is a major challenge. Emotion and sentiment annotation, in particular, is notoriously difficult and filled with results showing low agreement between annotators despite significant efforts by the researchers involved in the studies [7,11–14,20–23]. In part, this has been attributed to the subjectivity inherent in the task and to the choice of annotation schemas [9,10]. Early works in the field of emotion detection from text, based their annotation schemas on well-established emotion classification theories from the field of psychology, such as the six basic emotions of Ekman [8] and the eight of Plutchik's Wheel of Emotions [24]. This includes well-known benchmark datasets in the field, such as Tales [7], and ISEAR [25], all the way to UnifyEmotions [11] which was aggregated from 14 previously released datasets in emotion classification. More recently, researchers have begun to move away from these annotation schemas and release datasets with more emotion labels, a trend that has been naturally accompanied by a move away from multiclass annotation and into multi-label annotations. UnifyEmotions itself, despite being based on many multiclass datasets, attempted to move towards multi-label by recovering the information from original annotation files and ignoring disagreements. The relatively high profile of the SemEmval 2018 Affect in Tweets shared task (SemEval-2018 Task-1C) [12] also served to place more focus on multi-label emotion classification and on more fine-grained emotion labels. The shared task added optimism, pessimism, and love to Plutchik's 8 emotions for which they introduced a crowdsourced multi-label dataset with 11k tweets. Another interesting dataset is XED [26]. Although it uses Plutchik's 8 emotions as labels, the annotation was multi-label and contains 34k examples, drawn from subtitles. The more ambitious work in these directions was GoEmotions [13] which introduced the largest crowdsourced dataset to date, consisting of 58k Reddit comments annotated with 27 emotion labels. Our work goes further in this direction and introduces a multi-label dataset with 200k examples annotated with 38 emotions. Since it would be nearly impossible to crowdsource fine-grained emotions in literature, as we cannot reasonably expect crowd workers to produce good labels for the difficult-to-interpret language of literature, nor for the costs of such a task to be reasonable at scale, we instead opted for a semi-supervised approach. To summarize, the key differences are that our work uses a semi-supervised approach instead of expert annotators or crowdsource workers, introduces more fine-grained emotions (a total of 38), and is much larger (with 200k examples). Given that the application focus of our work is literature analysis, the best pre-existing alternative, Tales, was a single-label, small dataset (1.2k examples) annotated with only 5 emotion labels.

Multi-label fine-grained emotion detection from text is not just challenging for human annotators. It is also a challenge for machine learning algorithms. Similar to most tasks in Natural Language Processing, emotion detection approaches have moved toward the transformer architecture [27] which now dominates the state-of-the-art across the field (for examples of leaderboards in NLP, see https://gluebenchmark.com/leaderboard or https://huggingface.co/spaces/autoevaluate/leaderboards (accessed on 22 June 2023)). In XED, the authors used BERT [28] and reported a macro-averaged F1 score of 0.54 over

their 8 emotions plus neutral. In GoEmotions, the authors also used BERT and reported a macro-averaged F1 of 0.46 over their 27 emotions plus neutral. A recent evaluation of ChatGPT on this dataset reported only 26% of that [29]. On the SemEval 2018 dataset, recent work reports a 0.60 macro-averaged F1 using RoBERTa [30] over the 11 emotions plus neutral, which was an improvement compared to the previous best-reported results such as SpanEmo's 0.58 F1 [31]. We also choose RoBERTa as our model and report a very similar F1 score of 0.59 over our 38 emotions, which is impressive considering it is a semi-supervised dataset with more fine-grained labels than previous datasets.

The main method we use for building our semi-supervised dataset is ranking through a model trained on NLI. The viability of NLI based Zero-Shot classification to provide labels for emotion classification was established in Ref. [19], with model ensembles being more thoroughly covered in Ref. [32] and multiple templates in Ref. [33]. Both focused on the evaluation of NLI on existing emotion datasets, while this work focuses on creating a new dataset. We use two hypotheses for each emotion, where the main difference between them is the label rather than the template, which is similar to some prompts explored in Ref. [33]. Both of these previous works focused primarily on coarse-grained emotions, and evaluation of the Zero-shot classification, while the primary focus of ours is on fine-grained emotions and creating a new dataset. Zero-shot fine-grained emotion classification was done in Ref. [34] but only to serve as input to a sentiment classifier rather than focus directly on the emotion classification task. Related work has also shown that results of NLI based Zero-shot text classification can be improved in different ways by using ensembles of different hypothesis [33], ensembles of different models, and moving from zero-shot to few-shot [32]. Interestingly, fine-tuning pretrained NLI models for Zero-shot emotion classification can also be done without labeling data by using self-training [35]. In that work, the authors evaluate the entailment accuracy of their self-training approach on GoEmotions and conclude that it significantly improves results over the base NLI models. The key difference between our work and the related work is that rather than evaluate NLI Zero-Shot based approach on existing datasets, we use it to build a multi-label fine-grained dataset and evaluate fully supervised models trained on it.

Soft labeling data refers to assigning each instance a discrete probability, such as a likelihood or confidence, that it belongs to each class. In comparison, when using hard or crisp labels, instances are assigned to a class with a probability of 1 (a certain event in probability theory). Soft labels can be captured from human annotators [36] but are more common in semi-supervised approaches [37], where they are often a part of a self-training, co-training process, label propagation, or from derived a pseudo-labeling classifier. Soft labels can be converted to hard labels, usually by thresholding at a certain probability. Commonly, if an example has a probability greater than 0.5 of belonging to a class, it is assigned that class label when converting to hard labels. Soft labels are often credited with better results than hard labels, attributed to proving robustness to noise and implicit regularization (e.g., Refs. [38–41]). In this work, we will compare the use of soft and hard labels on our dataset.

The key differences between our work and previous works that introduced new multi-label emotion datasets [12,13,26] can be summarized as follows:

- Our work focuses on introducing a semi-supervised approach instead of crowdsourced workers for annotating training data;
- We introduce a dataset with more fine-grained emotions (38 labels) compared to previous datasets (11, 27, or 8 labels);
- Our work introduces the first multi-label or fine-grained emotion dataset for literature.

Compared to previous work that uses NLI based zero-shot emotion classification ([19,32,33]):

- We use NLI for binary ranking of candidates instead of directly providing labels;

- The final labels of our datasets are provided by a binary classifier for each emotion rather than by directly using NLI, allowing us to use only the highest NLI ranked examples for each emotion.

Finally, following previous work in semi-supervised learning, we will compare the use of soft labels in the context of this work.

1.4. Contribution

The purpose of this work is to improve the ability of researchers to analyze emotions in literature by providing a fine-grained emotion detection dataset and classifier. Further, we want to advance the state-of-the-art in emotion classification by providing a methodology that helps to develop emotion datasets. We tackle the issue of the existing emotion classification literature datasets being small and annotated with coarse, rather than fine-grained, labels. We overcome the challenge of how hard and expensive it is to annotate emotions in general, but harder in literature, by using a semi-supervised approach.

Our main contribution is a semi-supervised approach for creating multi-label emotion classification datasets, which we believe can be applied to similar problems, especially within the cultural heritage domain and other domains with a high cost of annotation. This approach can be used to create large balanced datasets without any human labeling.

Our second contribution is the dataset itself, with 38 fine-grained emotion labels, and its respective taxonomy and definitions that can be used for further research in emotion classification as well as for research related to improving the quality of semi-supervised data, such as methods for dealing with noisy labels. It includes more emotion labels than any other dataset, with GoEmotions [13] having only 27 and the only other literature sentence-level emotion detection dataset, Tales, only having 5 emotions. Similar to GoEmotions, but unlike Tales, it is also multi-label. Despite this, it still manages to have a balanced number of examples per label, which GoEmotions does not. We also provide an analysis of emotion correlation and emotion sentiment on our dataset that can serve to further inform future work in emotion detection and refine emotion taxonomies

Our final contribution is the model trained on our dataset, evaluated in multiple common scenarios and public benchmark datasets. The evaluation shows that the model trained on our dataset performs on par with models trained on crowdsourced datasets, and it provides good zero and few-shot transfer ability. Making the model public gives researchers in other fields, especially those analyzing literature, the necessary tools to perform emotion detection.

Our main contributions can be summarized as follows:

- A novel semi-supervised approach capable of creating fine-grained multi-label emotion classification datasets;
- A large, balanced dataset with 38 fine-grained emotion labels, surpassing existing datasets;
- A more comprehensive taxonomy and definitions for emotion detection from text;
- Analysis of emotion correlation and sentiment within the dataset, informing future work in emotion detection;
- Publicly available, easy-to-use trained models for researchers.

2. Materials and Methods

We start the description of materials and methods by describing Project Gutenberg, which contains a collection of free eBooks, from which we extract and filter sentences (Sections 2.1 and 2.2). Given our taxonomy (Section 2.3), we use a pretrained NLI model to perform by binary weak labeling (Section 2.4). We train models to do pseudo labeling and create the final multi-label emotion dataset (Section 2.5). These models, as well as subsequent models trained as part of our later experiments (Section 4), are detailed in Section 2.6.

2.1. Data

The biggest and most common source of open literature data is Project Gutenberg (PG). We use all English language books based on both metadata and language identification. The objective of the steps described in this section is to prepare the text for the next steps described in the following sections. First, since the objective is to perform sentence-level emotion detection, we must convert the books into sentences. Given that PG book files contains some data that is from the body of the book text (it may not be in English, or it may not be properly segmented into a sentence), we also apply some common heuristics to remove such potentially problematic sentences:

1. Strip away template text added to the book (https://github.com/c-w/gutenberg/ (accessed on 22 June 2023));
2. Check the language of the book based on the characters in the interval [1000:20,000];
3. Split the book into sentences based on the newline character first and then using a sentence tokenizer;
4. Discard any sentences in all uppercase (most likely headings or template text);
5. Discard any sentences that do not start with a character of the English alphabet;
6. Discard any sentences with less than 6 tokens or more than 40 (whitespace delimited);
7. Discard any sentences that are not identified as English or any sentences that contain 10-token segments that are not identified as English;
8. We randomly shuffled the sentences and limited their total number to 10 M to reduce the amount of compute required in the following sections.

To perform sentence segmentation, we used the NLTK [42] implementation of Punkt [43]. For Language identification, we used the fasttext language identification model [44,45].

2.2. Deduplication

We remove duplicates and near duplicates using several processes. The first is exact case-insensitive string matching, which removes exact duplicates. The second is MinHash Locality Sensitive Hashing [46] using the datasketch library (https://ekzhu.com/datasketch/ (accessed on 22 June 2023)) to estimate the Jaccard similarity between sentences. For pairs above the threshold of 0.9 we calculate the exact Jaccard similarity; if it is above this threshold we consider them duplicates. For this process, each sentence is represented as space-delimited token n-grams with n = (1, 2, 3) with text first being standardized by normalizing white spaces, removing control characters, removing accents, and converting it to lower case.

2.3. Emotion Taxonomy

We developed a taxonomy of 38 fine-grained emotions, plus neutral, based primarily on existing work in NLP as well as work in other fields. The emotion labels and their respective definitions are listed in Table 1. The definitions are based on dictionary definitions and synonyms, primarily, the Oxford English Dictionary (https://www.oed.com/ (accessed on 22 June 2023)) and the Merriam-Webster dictionary (https://www.merriam-webster.com/ (acessed on 22 June 2023)).

Our taxonomy directly includes all the Ekman basic emotions. From the Plutchik basic emotions, "Anticipation" is the only one directly missing, as it a compound of very different emotions we include such as "optimism" and "boredom". With regard to the GoEmotions taxonomy, 26 of the 27 emotions are present, with only "realization" missing. We found it difficult to provide a meaningful non-trivial, unambiguous definition that would separate "realization" from "surprise" and be clear for potential human annotators. The emotions "desire", "love", and "nostalgia" are included based on requirements of historians involved in our project. Additionally, we spent some time looking for other emotions in the Wikipedia page for emotion classification (https://en.wikipedia.org/wiki/Emotion_classification (accessed on 22 June 2023)). After exploratory experiments, we settled on those described in Table 1, however, we see the potential for adding more in the future.

Table 1. Emotion Taxonomy.

Emotion	Definition
admiration	finds something admirable, impressive or worthy of respect
amusement	finds something funny, entertaining or amusing
anger	is angry, furious, or strongly displeased; displays ire, rage, or wrath
annoyance	is annoyed or irritated
approval	expresses a favorable opinion, approves, endorses or agrees with something or someone
boredom	feels bored, uninterested, monotony, tedium
calmness	is calm, serene, free from agitation or disturbance, experiences emotional tranquility
caring	cares about the well-being of someone else, feels sympathy, compassion, affectionate concern towards someone, displays kindness or generosity
courage	feels courage or the ability to do something that frightens one, displays fearlessness or bravery
curiosity	is interested, curious, or has strong desire to learn something
desire	has a desire or ambition, wants something, wishes for something to happen
despair	feels despair, helpless, powerless, loss or absence of hope, desperation, despondency
disappointment	feels sadness or displeasure caused by the non-fulfillment of hopes or expectations, being or let down, expresses regret due to the unfavorable outcome of a decision
disapproval	expresses an unfavorable opinion, disagrees or disapproves of something or someone
disgust	feels disgust, revulsion, finds something or someone unpleasant, offensive or hateful
doubt	has doubt or is uncertain about something, bewildered, confused, or shows lack of understanding
embarrassment	feels embarrassed, awkward, self-conscious, shame, or humiliation
envy	is covetous, feels envy or jealousy; begrudges or resents someone for their achievements, possessions, or qualities
excitement	feels excitement or great enthusiasm and eagerness
faith	expresses religious faith, has a strong belief in the doctrines of a religion, or trust in god
fear	is afraid or scared due to a threat, danger, or harm
frustration	feels frustrated: upset or annoyed because of inability to change or achieve something
gratitude	is thankful or grateful for something
greed	is greedy, rapacious, avaricious, or has selfish desire to acquire or possess more than what one needs
grief	feels grief or intense sorrow, or grieves for someone who has died
guilt	feels guilt, remorse, or regret to have committed wrong or failed in an obligation
indifference	is uncaring, unsympathetic, uncharitable, or callous, shows indifference, lack of concern, coldness towards someone
joy	is happy, feels joy, great pleasure, elation, satisfaction, contentment, or delight
love	feels love, strong affection, passion, or deep romantic attachment for someone
nervousness	feels nervous, anxious, worried, uneasy, apprehensive, stressed, troubled or tense
nostalgia	feels nostalgia, longing or wistful affection for the past, something lost, or for a period in one's life, feels homesickness, a longing for one's home, city, or country while being away; longing for a familiar place
optimism	feels optimism or hope, is hopeful or confident about the future, that something good may happen, or the success of something
pain	feels physical pain or is experiences physical suffering
pride	is proud, feels pride from one's own achievements, self–fulfillment, or from the achievements of those with whom one is closely associated, or from qualities or possessions that are widely admired
relief	feels relaxed, relief from tension or anxiety
sadness	feels sadness, sorrow, unhappiness, depression, dejection
surprise	is surprised, astonished or shocked by something unexpected
trust	trusts or has confidence in someone, or believes that someone is good, honest, or reliable

2.4. Weak-Labeling

An overview of the weak labeling process is shown in Figure 1. We assign weak labels to example sentences using Zero-Shot classification with BART [47] pretrained for NLI (https://huggingface.co/facebook/bart-large-mnli (accessed on 22 June 2023)). Textual entailment is a classification task where, given a pair of texts, one being a hypothesis and the other a premise, the objective of the classifier is to decide whether the hypothesis is true given the premise. More specifically, in the MNLI [48] dataset which was used to train the off-the-shelf NLI model used in this work, it is formulated as a 3-class multiclass problem: given a premise-hypothesis pair, the classifier must decide whether they are in entailment, in contradiction, or are neutral in relation to each other. Using the target sentence as the premise and creating a sentence expressing the hypothesis that the premise expresses a given emotion, the probability of entailment becomes the probability that the emotion is expressed in the target sentence. To create this probability, we use the unnormalized outputs from the network, corresponding to "entailment" and "contradiction" and pass them through the Softmax.

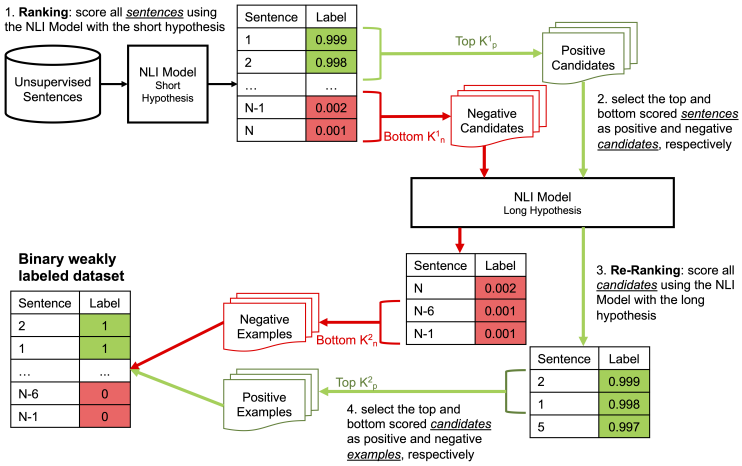


Figure 1. Weak labeling process using an existing NLI model to identify positive and negative examples of sentences express emotions. The key steps, Ranking and Re-Ranking, in bold, use different templates described in this section. Inputs and outputs of each step underlined: from sentences, to candidates, to examples.

Weak labels are assigned in a two-stage process for each emotion. First, we use the NLI model with the template *This expresses the emotion {emotion}* to rank all examples obtained in Section 2.2 by the probability of entailment. Second, we load the top 150k examples by this score and re-rank them using the NLI model with the template *Speaker or someone {emotion definition}*. Table 2 shows our ranking and re-ranking hypothesis resulting from applying our templates to an emotion described in Table 1. From here, we take the top 6000 sentences by their re-ranked order and assign to them the respective emotion label.

Table 2. Example of NLI hypothesis for the emotion "fear". See also Table 1.

	Hypothesis
Short (Ranking)	This expresses the emotion fear.
Long (Re-ranking)	Speaker or someone is afraid or scared due to a threat, danger, or harm.

Next, we filter the examples to ensure semantic diversity. Using semantic embeddings provided by a MiniLM [49] model (https://huggingface.co/nreimers/MiniLM-L6-H384-uncased (accessed on 23 June 2023)) pretrained on multiple semantic similarity tasks [50]. Examples with a cosine similarity greater than 0.7 were discarded. Finally, we filter examples using simple token and stem [51] counting. The values were manually tweaked per emotion, based on the output. The max stem count was set between 600 while the maximum token count varied between 1000 and 2000. This variation is to account for the variation in the vocabulary for expressing different emotions. For example, "boredom" has a much more narrow vocabulary than "disgust". The top 6000 examples by score are selected as positives examples for the emotion.

The same process is repeated for selecting negative examples except we pick the lowest scored examples and take the bottom 300,000 examples by ranking score, followed by the bottom 10,000 examples by the re-ranking score. A separate process handles neutral (no emotion) examples. First, the sentences are filtered by the NLI model using the hypothesis *This text expresses emotion*. Only sentences with a probability below 0.2 are selected. These are then filtered again by the NLI model using the hypothesis *The speaker or someone expresses or feels {emotion}* for each emotion in our taxonomy. Only sentences with a probability below 0.5 for all emotions are selected as neutral examples. Approximately 5100 neutral examples were created in this way. At the end of this process described, we have a dataset with binary labels for each emotion, totaling 38 datasets, each with 26,000 examples, plus the neutral examples.

2.5. Pseudo-Labeling

In this section, we create our final dataset. We do this by converting the emotion-specific binary datasets created by weak-labeling in Section 2.4 into a multi-label dataset by pseudo-labeling [52] the data. For each emotion, we use the weakly labeled examples to train a binary classifier for that emotion. This classifier is then used to classify all examples in our dataset except the neutral examples (these are already multi-label by definition). Figure 2 shows a diagram of this process. The classifier is described in Section 2.6 and is trained with a label smoothing factor [53] of 0.2 with the other hyperparameters detailed in Table 3.

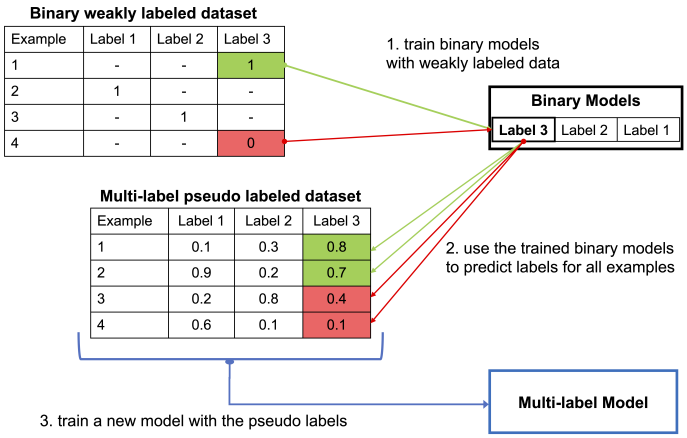


Figure 2. Pseudo labeling using weakly labeled data to train a model. Positives in green, negatives in red.

Table 3. Hyperparameter values used for training the binary classifiers.

Hyperparameter	Value
Batch Size	32
Learning Rate	3×10^{-5}
Max Epochs	2
Smoothing	0.2

The output of each binary classifier is a Softmax score that can be interpreted as the probability that the sentence expresses that particular emotion. We explore the use of this score as a soft label in Section 4. Applying the threshold of $\geq$0.5 to these scores can be used to create hard labels (0 or 1).

To make the dataset more manageable, we reduce its size to 200K examples by sub-sampling it using stratified splitting (https://github.com/trent-b/iterative-stratification (accessed on 23 June 2023)) [54] and further split it into train, validation, and test sets with a further small sample set aside for manual annotation. The size is still more than large enough to allow for further pruning or subsampling, while not being so big as to hinder experimentation. Pre-splitting into different subsets allows for easier comparison of any future works based on this dataset.

2.6. Supervised Classification

In this work, we train two sets of supervised text classifiers. The first set consists of the binary emotion-specific classifiers that perform the pseudo-labeling process described in Section 2.5. The second set consists of the classifiers used in the experiments described in Section 4. Both sets follow the same architecture as shown in Figure 3. This is a pre-trained language model based on the transformer architecture [27] with a classification head for task-specific fine-tuning, which has come to dominate the state of the art in NLP. We use RoBERTa [55] as the encoder, which replicates BERT [28] with better training. It also replaces BERT's character-level Byte-Pair Encoding (BPE) [56] with a byte-level implementation

similar to GPT [57]. This step is commonly called tokenization, as it segments text into a list of integer identifiers mapped to embeddings, in the model's embedding layer. We do not use additional preprocessing or text encoding steps. The classification head consists of dropout [58], followed by a fully connected dense layer, a hyperbolic tangent activation, dropout, and a fully connected dense projection layer. This classification head is used by the Transformers library in the *"RobertaForSequenceClassification"* model. We adopt it as it allows for easier replication of our work and for our model weights to be shared without the need to share additional code (See https://huggingface.co/lrei/emolit-roberta-large-soft for an example (accessed on 23 June 2023).

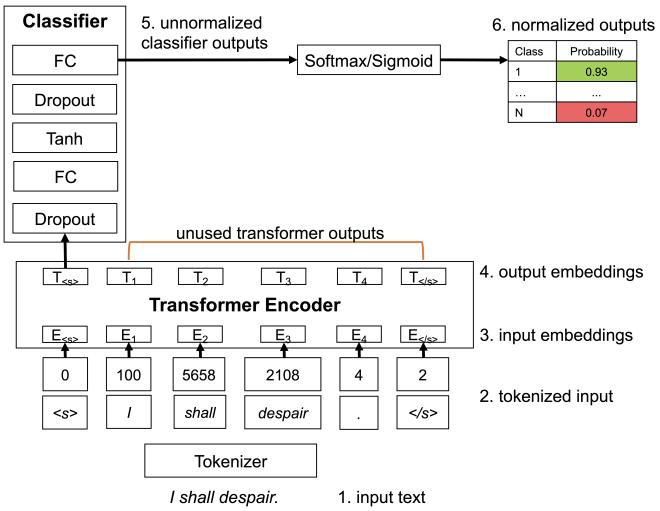


Figure 3. Architecture of the Classifier model. The main blocks are the transformer encoder and the classifier, in bold. Only the output of the start token is passed to the classifier, all other encoder outputs are unused (in orange). Example classified positive label in green, negative label in red.

The pseudo-labeling models use the smaller *BASE* transformer with 12 transformer layers, a hidden dimension of 768, with 12 attention heads (L = 12, H = 768, A = 12), totaling 125 M parameters. The models in the experiments section primarily use the *LARGE* transformer (L = 24, H = 1024-hidden, A = 16, 355 M parameters). Additionally, in Section 4.2 we also provide results for a smaller and faster model based on the *DISTIL* (https://huggingface.co/distilroberta-base (accessed on 23 June 2023)) transformer (L = 6, H = 768, A = 12, 82 M parameters) [59].

Pseudo-labeling models use the Softmax as the final activation function, whereas the models in the experiments use either the Softmax or Sigmoid based on whether the target dataset is multiclass or multi-label, respectively. The final FC layer has a size of $H \times N$ where N is the number of classes, for pseudo-labeling N = 2, for models trained on our data N = 38, while models trained on other datasets have dataset-specific values of N. The dropout probability used was always 0.1.

All models are trained using mixed precision (FP16) and the Adam [60] optimizer with the linear scheduler without warmup. Hyperparameters are taken directly from Ref. [55] without any hyperparameter tuning specific to the dataset. For pseudo-labeling, we used a batch size of 32 with a learning rate of 3×10^{-5}. Experiment-specific hyperparameters are presented in their respective sections. All models trained on our dataset were limited during training to a maximum sequence length of 48. The computational complexity of the transformer architecture, resulting from the use of self-attention, is $\mathcal{O}(L \cdot (n^2 \cdot H))$, where L is the number of self-attention layers, n is the sequence length, and H is the hidden dimension size. A neural network's memory requirements are defined by the number of parameters it has, and how they are stored. The *LARGE* models have 335 M million parameters, and we use 2 bytes per parameter (with FP16 inference), which results in approximately 650 MB of memory used by the model. With 160k training examples and the

hyperparameters used, the *LARGE* model takes about 3 h to train on an NVIDIA GeForce RTX 3090.

2.7. Evaluation Metrics

Throughout Section 4, we will report experimental results using the macro-averaged F1 metric (also known as maF1) as a single number summary. This is the standard metric for reporting multi-label classification results in the field (see Section 1.3), especially with imbalanced test sets. In this section, we provide a brief explanation of this metric and the other metrics used: precision and recall.

Multi-label classification is a type of classification problem where each instance can be assigned to multiple classes simultaneously. In contrast to binary classification, where an instance is assigned to either one of two classes, multiclass classification, where each instance is assigned to exactly one class out of a set of mutually exclusive classes. In multi-label classification, each class label can be independently predicted, and an instance can belong to none, one, or multiple labels. In the context of multi-label classification, the terms True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN) are defined based on the prediction and ground truth of each class separately.

- True Positives (TP): For a specific label, TP represents the instances where the model correctly predicts the presence of that label, and the ground truth also indicates the presence of that label. It is the count of instances that are correctly identified as positive for a particular label;
- False Positives (FP): For a specific label, FP occurs when the model predicts the presence of that label, but the ground truth indicates the absence of that label. It is the count of instances that are incorrectly classified as positive for a particular label;
- True Negatives (TN): For a specific label, TN represents the instances where the model correctly predicts the absence of that label, and the ground truth also indicates the absence of that label. It is the count of instances that are correctly identified as negative for a particular label;
- False Negatives (FN): For a specific label, FN happens when the model predicts the absence of that class, but the ground truth indicates the presence of that label. It is the count of instances that are incorrectly classified as negative for a particular label.

Rather than directly reporting these raw counts, it is common to report precision, recall, and F1, which provide a more comprehensive measure of the model's performance as they take into account proportions rather than raw counts.

- Precision measures the proportion of true positive predictions out of the total positive predictions (Equation (1)). It focuses on the correctness of positive predictions. A high precision indicates a low rate of false positives;
- Recall measures the proportion of true positive predictions out of the total positive instances in the dataset (Equation (2)). It focuses on capturing all positive instances without missing any. A high recall indicates a low rate of false negatives;
- F1 score is the harmonic mean of precision and recall. It provides a balanced measure that combines both precision and recall into a single metric (Equation (3)). It is commonly used when both precision and recall are equally important, providing an overall measure of the model's performance.

$$precision = \frac{TP}{TP + FP} \tag{1}$$

$$recall = \frac{TP}{TP + FN} \tag{2}$$

$$F_1 = 2\frac{precision \cdot recall}{precision + recall} \tag{3}$$

Accuracy is calculated as the ratio of the total number of correct predictions (both true positives and true negatives) to the total number of instances in the dataset (Equation (4)). In datasets, where the number of negative instances is much higher than the number of positive instances, the accuracy is usually misleading. A classifier that always predicts the majority class (negative) would achieve high accuracy due to the imbalance, even if it fails to correctly identify positive instances. In multi-label datasets with many labels, it is common for every label to have many more negatives than positives. Considering our problem, with 38 emotion labels, a given sentence is unlikely to express any one given emotion. This will be shown in Section 3), where we will see that most examples have at most two emotion labels. Thus, the ratio of positives to negatives in any test set is highly imbalanced; e.g., less than 2% of examples are positives in multiple classes and no class has more than 18% positives in our gold set. The accuracy score will be entirely dominated by negatives, with classes having high accuracy based of a good true negative rate.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

In multi-label classification tasks, where there are multiple labels to predict, the macro-average is often used to calculate metrics such as precision, recall, and F1 score. Macro-average calculates the metric separately for each label and then takes the average across all labels. It treats each label equally, regardless of the label's frequency or imbalance in the dataset. This approach ensures that each label's performance contributes equally to the overall metric. The formula for the macro-average F1 score (Equation (5)) is calculated by a simple average of class-specific F1 scores (F_{1_i}) over all classes (C) in the test set [61].

$$maF_1 = \frac{1}{C}\sum_{i=1}^{C} F_{1_i} \tag{5}$$

This is the primary measure used in the literature and, sometimes, the only one reported (e.g., Ref. [26]), and it is thus necessary for comparison with existing work. We will also report precision and recall, both per class and macro-averages.

3. Data Analysis

The purpose of this section is to provide an understanding of one of the key results of our work: the emotion detection dataset we built. To do so, we provide its basic statistics, namely, the total number of sentences, how many sentences per emotion, and how many emotions per sentence (Section 3.1). We also provide a brief analysis of the correlation between emotions (Section 3.2) and summarize the sentiment associated with each emotion (Section 3.3).

3.1. Label Distribution

Label counts for our dataset are shown in Table 4. Datasets created from randomly sampled data tend to have a long-tailed label distribution. This is generally more pronounced when fine-grained labels are used. For example, in GoEmotions, the most common emotion label is "admiration" which has 53 times (53.64 times, measured on the train set) as many examples as the least common emotion label, "grief", despite the use of a "pilot model" to balance emotions. In our case, the most common label, "nostalgia" (Appendix C explains why), is only 2 times more common than the least common label, "greed". The balanced distribution and the large size of our dataset (200k sentences) should allow for the use of automatic pruning methods to reduce noise or sampling for the creation of manually labeled datasets.

Table 4. The number of examples per emotion for each data subset, including the human-labeled gold dataset. We nearly achieved a balanced distribution of examples per emotion label.

	Train	Validation	Test	Gold
admiration	7700	963	962	110
amusement	7427	928	928	52
anger	8556	1070	1070	72
annoyance	10,730	1341	1341	57
approval	8531	1066	1066	98
boredom	8113	1014	1014	53
calmness	8573	1072	1072	45
caring	8972	1122	1121	64
courage	8484	1061	1060	42
curiosity	7738	967	967	67
desire	10,160	1270	1270	81
despair	10,009	1251	1251	44
disappointment	12,133	1517	1517	39
disapproval	11,130	1391	1391	111
disgust	8987	1123	1123	72
doubt	9012	1127	1127	43
embarrassment	9642	1205	1205	22
envy	9942	1243	1243	14
excitement	10,794	1349	1349	38
faith	8442	1055	1055	13
fear	11,556	1445	1445	39
frustration	11,162	1395	1395	54
gratitude	11,279	1410	1410	14
greed	7423	928	928	25
grief	10,972	1372	1371	14
guilt	8660	1082	1082	13
indifference	8549	1069	1069	37
joy	9404	1175	1175	61
love	8838	1105	1105	50
nervousness	7747	968	968	24
nostalgia	14,805	1851	1851	29
optimism	9560	1195	1195	37
pain	10,014	1252	1252	22
pride	10,744	1343	1343	27
relief	9317	1165	1165	25
sadness	9589	1199	1199	52
surprise	9818	1227	1227	36
trust	8606	1076	1076	43
neutral	22,803	2890	2919	15
Sentences	160,000	20,000	20,000	727

Table 5 shows the number of labels per example as a percentage. More than 60% of the examples have between 0 and 2 emotion labels, while only around 13% have 5 or more emotion labels. For comparison, in GoEmotions, 83% of the examples have only 1 label. This is because the authors of GoEmotions specifically optimized their taxonomy to minimize the number of commonly overlapping emotions and limit the number of emotions. In our work, we instead tried to have the highest number of emotions that we could, for a first attempt, while still being able to clearly define and separate them. This more fine-grained approach to the taxonomy is designed to enable more precise study. For example, "despair" might be very similar to "sadness" but its difference can be very significant, not just when analyzing fiction but also non-fiction. One example is medical texts, where "despair" can be more closely related to depression than "sadness". Section 3.2 deals with emotion correlation.

Table 5. Number of emotions per example (%). The majority of sentences have two or fewer emotion labels.

No of Emotions	Examples (%)
0	14.3
1	27.4
2	20.8
3	14.4
4	10.0
5	6.6
6	4.3
7	2.3

3.2. Label Correlation

We calculated the correlation between each emotion over the sentences in our dataset, that is, emotions that occur in the same sentence (since sentences can express multiple emotions given a multi-label formulation), and present the results in Figure 4. The highest and lowest label pairs are summarized in Table 6. The results are mostly common sense, given the definitions, and are extremely similar to those in GoEmotions, with the top pairs being the same where possible (e.g., annoyance and anger, joy and excitement, nervousness, and fear). Very low negative correlation between emotions is likely less common in literature compared to social media due to the purposeful use of contrast. For example, it is common to praise one character while in the same sentence showing contempt for others ("I liked your spirit; I hate these tame, perfectly conventional girls; they bore me to death"). Or show hope in the middle of despair ("I feel at times very much deprest indeed, but have no doubt but that all things will work for good").

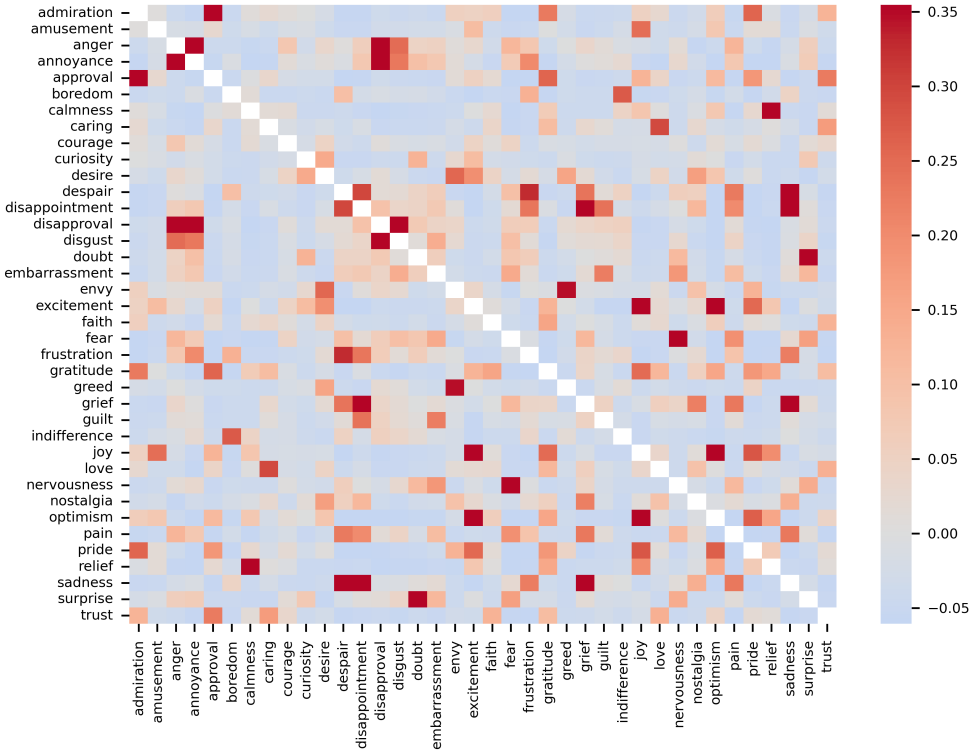


Figure 4. Emotion correlation calculated over the sentences in our dataset. Mostly common sense, except for smaller values of negative correlation. The results are in line with other datasets.

Table 6. Summary of emotion correlation: highest and lowest correlated emotions. The results are very similar to GoEmotions.

Highest Correlation		**Lowest Correlation**	
despair	sadness	optimism	pain
calmness	relief	annoyance	optimism
fear	nervousness	approval	frustration
anger	annoyance	frustration	gratitude
excitement	joy	disappointment	optimism

3.3. Label Sentiment

We performed sentence-level sentiment analysis on our dataset by using an off-the-shelf sentiment model (https://huggingface.co/distilbert-base-uncased-finetuned-sst-2-english (accessed on 23 June 2023)). For each emotion, we count every positive and negative sentence labeled with that emotion. The resulting ratio is shown in Figure 5 with a summary of the most positive and negative emotions in Table 7.

Figure 5. Sentiment of emotions: ratio of positive and negative sentiment sentences labeled with each emotion in our dataset.

Table 7. Summary of emotion sentiment: most positive and most negative emotions. Most positive and most negative emotions match common sense expectations.

Most Positive		**Most Negative**	
Emotion	**Positive (%)**	**Emotion**	**Negative (%)**
admiration	92	frustration	88
approval	91	boredom	86
optimism	90	despair	86
trust	89	annoyance	83
joy	89	pain	82

Again, the results are mostly common sense. Where they are immediately obvious, it is easier to understand in conjunction with the label correlations in Figure 4 and by keeping in mind the definitions in Table 1. In Section 3.2 we discussed the use of contrast results in low negative correlation for emotions that should be opposites by common sense reasoning. In part, this helps explain why the sentiment of emotions is generally not more extreme

(i.e., closer to 100% negative or positive). The emotion that jumps to our attention is "envy". At first, by common sense, it seems like it ought to be a more negative emotion, but the relatively high correlation with "admiration" helps explain why it is not: admirable things are enviable, but that is not necessarily negative. For example, "Even the pattern from which she is working, the silk, the gold, the lawn, made happy by her touch, are sanctified, are envied".

4. Experimental Results

4.1. Evaluation Data

We use four well-established, public, emotion benchmark datasets. Tales is the most relevant public benchmark dataset for our work. It is an affect dataset with 1207 sentences built from literature, namely fairy tales by B. Potter, H.C. Andersen, and the Brothers Grimm [7]. It was annotated with the six basic Ekman emotions; however, anger and disgust were merged into a single label. GoEmotions [13] is the second most relevant public benchmark dataset for our work. It is a fine-grained emotion dataset consisting of 54,260 Reddit posts annotated by crowd workers with 27 emotions plus a neutral label. ISEAR (International Survey on Emotion Antecedents and Reactions) [25] is a collection of 7666 sentences each answering a question assigned to 7 emotions: joy, fear, anger, sadness, disgust, shame, and guilt. EMOINT [62] contains crowdsourced annotations for 7102 tweets annotated with anger, joy, sadness, and fear. The Tales dataset is relevant because it built from literature. GoEmotions is relevant because it is annotated with fine-grained emotions. We included ISEAR and EMOINT as benchmark datasets in these experiments because they are commonly used in related work. Both are also part of UnifyEmotions [11] and are thus easily accessible.

In addition to the public benchmark datasets, we also manually annotated a subset of our dataset with our own taxonomy. This is a set of 727 sentences annotated with all 38 emotions and detailed in Table 4 under the header "Gold". It was created by iterative stratified sampling from the original dataset obtained in Section 2.5. A total of 1000 sentences were originally extracted, but to reduce the annotation work, 100 expected neutrals were discarded, and ultimately only the first 727 examples were manually labeled (labeling was performed by the first author).

4.2. Supervised Evaluation

Our first experiment consists of training two models on the train split of our dataset and evaluating them on our labeled gold split. The model architecture is described in Section 2.6. The training was done for 10 epochs, with the best epoch on the validation split being used to provide the final results. This setting is based on the original RoBERTa paper [55]. These models are effectively a baseline designed to evaluate our dataset.

The models differ only in the fact that one is trained on hard labels (0 or 1) created from applying the 0.5 threshold on the soft labels that the other model is trained on. The soft labels are the probability output of the pseudo-labeling classifiers described in Section 2.5. The results are shown in Table 8 with the corresponding hyperparameters shown in Table 9.

The macro averaged scores for both models are essentially the same at 0.59 F1. The results appear to be in line with what a similar model would achieve when trained and evaluated on GoEmotions, which is annotated by crowd workers while our dataset was automatically annotated. This is a good outcome. We expected the use of soft labels to provide some regularization and improved results on the human-labeled data, but this was not the case. Additionally, we compare the results with different model sizes without changing the architecture. Table 10 shows the summary comparison of different encoder sizes. There was no significant difference between using the BASE encoder and the LARGE encoder when fine-tuned and evaluated on our dataset with similar hyperparameters. The DISTIL model, which uses the distilroberta-base encoder, had a slightly worse F1 score. When comparing our RoBERTa classifier with the more common BERT baseline, we see that RoBERTa outperforms BERT in both the DISTIL and BASE sizes. The results for the

LARGE size are the same. We can also see that RoBERTa LARGE and BASE have the same score but are still better than the DISTIL size. This is a data point that indicates likely diminishing returns from increasing the model size and suggests that future work should focus on improving the dataset.

Table 8. Results on the human-annotated portion of the EmoLit dataset. The soft and hard label models had the same average F1 score. The results are in line with related work (e.g., GoEmotions).

	Hard Labels			**Soft Labels**		
Emotion	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
admiration	0.74	0.31	0.45	0.72	0.31	0.43
amusement	0.73	0.87	0.79	0.75	0.87	0.8
anger	0.70	0.65	0.68	0.71	0.68	0.69
annoyance	0.51	0.74	0.60	0.5	0.72	0.59
approval	0.83	0.51	0.63	0.8	0.49	0.61
boredom	0.67	0.94	0.78	0.64	0.92	0.75
calmness	0.65	0.82	0.73	0.63	0.82	0.71
caring	0.73	0.83	0.77	0.69	0.8	0.74
courage	0.47	0.67	0.55	0.46	0.67	0.54
curiosity	0.76	0.82	0.79	0.76	0.84	0.79
desire	0.82	0.79	0.81	0.81	0.77	0.78
despair	0.72	0.70	0.71	0.7	0.7	0.7
disappointment	0.44	0.46	0.45	0.4	0.44	0.42
disapproval	0.47	0.23	0.31	0.48	0.25	0.33
disgust	0.84	0.38	0.52	0.79	0.36	0.5
doubt	0.72	0.49	0.58	0.61	0.44	0.51
embarrassment	0.58	0.64	0.61	0.5	0.73	0.59
envy	0.28	0.86	0.42	0.28	0.93	0.43
excitement	0.55	0.68	0.61	0.57	0.68	0.62
faith	0.39	0.85	0.54	0.45	0.77	0.57
fear	0.48	0.41	0.44	0.42	0.41	0.42
frustration	0.54	0.57	0.56	0.53	0.61	0.57
gratitude	0.28	0.79	0.42	0.26	0.71	0.38
greed	0.57	0.64	0.60	0.55	0.68	0.61
grief	0.27	0.86	0.41	0.31	0.93	0.46
guilt	0.43	0.69	0.53	0.45	0.77	0.57
indifference	0.66	0.89	0.76	0.65	0.84	0.73
joy	0.84	0.43	0.57	0.77	0.44	0.56
love	0.69	0.66	0.67	0.69	0.72	0.71
nervousness	0.54	0.54	0.54	0.55	0.46	0.5
nostalgia	0.29	0.97	0.44	0.27	0.97	0.42
optimism	0.52	0.43	0.47	0.5	0.38	0.43
pain	0.33	0.55	0.41	0.42	0.73	0.53
pride	0.46	0.59	0.52	0.48	0.59	0.53
relief	0.54	0.84	0.66	0.51	0.84	0.64
sadness	0.70	0.60	0.65	0.67	0.62	0.64
surprise	0.68	0.69	0.68	0.71	0.69	0.70
trust	0.71	0.63	0.67	0.76	0.65	0.70
macro-average	0.58	0.66	0.59	0.58	0.66	0.59
std	0.17	0.18	0.13	0.16	0.18	0.13

Table 9. Hyperparameter values used for training the supervised model on our dataset.

Hyperparameter	**Value**
Batch Size	16
Learning Rate	2×10^{-5}
Max Epochs	10

Table 10. Comparison between different model sizes trained on our training set and evaluated on the gold set. The F1 score is similar for all RoBERTa model sizes, although DISTIL is slightly worse (0.59 vs. 0.58). BERT underperforms for RoBERTa except for LARGE.

Encoder Name	Encoder Architecture	Encoder Parameters	F1 (Macro)
RoBERTa-large	L = 24, H = 1024, A = 16	355 M	0.59
BERT-large	L = 24, H = 1024, A = 16	340 M	0.59
RoBERTa-base	L = 12, H = 768, A = 12	125 M	0.59
BERT-base	L = 12, H = 768, A = 12	110 M	0.58
DistilRoBERTa-base	L = 6, H = 768, A = 12	82 M	0.58
DistilBERT-base	L = 6, H = 768, A = 12	66 M	0.56

Finally, in Table 11, we compare with the results provided by related work, showing that the baseline results in our dataset, created by our semi-supervised approach, are similar to crowdsourced multi-label emotion datasets. This is true even though our taxonomy is more fine-grained and thus needs to distinguish between very similar emotions that are grouped together in the other datasets.

Table 11. Comparison with the results reported in related work, using *BASE* model sizes.

Dataset	Emotions	Model	F1 (Macro)
EmoLit (ours)	38	RoBERTa	0.59
		BERT	0.58
SemEval-2018 Task-1C	11	RoBERTa	0.60 [30]
XED	8	BERT	0.54 [26]
GoEmotions	27	BERT	0.46 [13]

4.3. Zero-Shot Transfer

Our taxonomy (Table 1) includes all emotions used in the selected public benchmark datasets, except for "realization" in GoEmotions. As such, it is possible to directly evaluate a model trained on our data on these datasets without any adaptation by simply mapping labels (See Appendix A). To handle "realization", we simply merge it with "surprise" (based on the definition and mapping provided in Ref. [13]). This creates the dataset we refer to in this section as "GoEmotions26". This experimental setup is, conceptually, the inverse of the UnifyEmotions [11] where instead the labels of the target dataset are mapped into a common taxonomy and the classifiers are trained and evaluated on that. Such an evaluation makes the datasets and classifiers use a more coarse-grained taxonomy than originally intended.

To compare the results, we train and evaluate models using 5-fold cross validation on Tales, ISEAR, and EMOINT. These models follow the same architecture described in Section 2.6 and are trained with the hyperparameters in Table 12. We also trained a model on GoEmotions, using the same model architecture, following the exact same procedure as the supervised model trained on EmoLit (Section 4.2) and the same hyperparameters (Table 9 with the model from the best epoch selected on the basis of the GoEmotions development set. Table 13 shows the results.

Table 12. Hyperparameter values used for training the Cross-Validation models.

Hyperparameter	Value
Batch Size	8
Learning Rate	2×10^{-5}
Epochs	3

Table 13. Zero-shot transfer experiment macro F1 on benchmark datasets and relative percentage compared to training on the same dataset (Self). The best transfer results correspond to a model trained on our dataset with soft labels.

Dataset	Tales	ISEAR	EMOINT	GoEmotions26
Self	0.83	0.76	0.82	0.52 [1]
Transfer				
GoEmotions	0.74 (89%)	0.52 (68%)	0.50 (61%)	NA [2]
EmoLit (Hard Labels)	0.74 (89%)	0.53 (70%)	0.57 (70%)	0.28 (54%)
EmoLit (Soft Labels)	0.77 (93%)	0.56 (74%)	0.60 (73%)	0.28 (54%)

[1] The model was trained on the 27 emotion train split of this dataset, mapping to 26 emotions done on the output. [2] Zero-Shot evaluation is not possible.

In this experiment, training the model on soft labels outperformed hard labels. Both models trained on our dataset outperformed the model trained on GoEmotions on the other datasets, with the best being the model trained with soft labels, which attains an average score of 0.62 over the first 3 datasets, followed by the model trained on hard labels at 0.61 and lastly, the model trained on GoEmotions with 0.59. Unsurprisingly, training models on the same dataset (relative to evaluation) significantly outperform zero-shot transfer. The conclusion of this experiment is that models trained on our dataset learn about the emotions they are trained to distinguish and can be used in zero-shot emotion classification. Surprisingly, the model trained on GoEmotions (Reddit posts) did not outperform models trained on our dataset (literature) when evaluated on EMOINT (tweets) but it did match our results on Tales (literature).

4.4. Few-Shot Transfer

In this experiment, we evaluate the models trained on our dataset in a common transfer learning scenario. This largely mirrors the experimental setup in Ref. [13]: we replace the classification head with a newly initialized classification head for each benchmark dataset and fine-tune the models on a sample of the data. The amount of training data used from each target dataset is 50, 100, 150, and 200 examples, and sampling was stratified. Evaluation is conducted on the remaining data in the benchmark dataset with results shown in Table 14. The baseline is the same model architecture (see Section 2.6) fine-tuned on the few-shot examples. The comparison model was trained on the GoEmotions dataset (the same model from Section 4.3). All fine-tuning in these experiments used a batch size of 8 over 8 epochs with a learning rate of 2×10^{-5} (summarized in Table 15). The results are an average of three runs.

Table 14. Fine-tuning transfer learning on benchmark datasets (macro F1). Transfer models outperform the baseline at a low number of examples. The best transfer results depend on the target dataset.

	Tales Literature				ISEAR Self-Reporting				EMOINT Tweets			
	50	100	150	200	50	100	150	200	50	100	150	200
Baseline	0.2	0.57	0.64	0.82	0.17	0.43	0.25	0.62	0.17	0.21	0.47	0.72
GoEmotions	0.58	0.75	0.80	0.81	0.47	0.58	0.62	0.63	0.53	0.62	0.67	0.68
EmoLit (Hard Labels)	0.60	0.72	0.79	0.81	0.36	0.52	0.55	0.60	0.56	0.65	0.69	0.71
EmoLit (Soft Labels)	0.61	0.74	0.79	0.81	0.36	0.54	0.55	0.62	0.59	0.67	0.70	0.72

Table 15. Hyperparameter values used for few-shot finetuning.

Hyperparameter	Value
Batch Size	8
Learning Rate	2×10^{-5}
Epochs	8

All models trained on emotions significantly outperform the baseline at a low number of examples. As the number of examples increases, the difference between fine-tuning a model pretrained for emotion classification and just the base model pretrained on the self-supervised language modelling objective decreases. At 200 examples, they are either identical or near-identical. The actual results on each benchmark are very similar for all models pretrained on emotion datasets—ours or GoEmotions. We can conclude that our dataset performs equally well in fine-tuning based few-shot transfer learning. Zero-shot results (Table 13 are significantly better for Tales and ISEAR, indicating that adaptation based on not discarding the classification head would likely be preferable. Zero-shot would also be preferable for both datasets if less than 200 examples were available for Tales, less than 100 for ISEAR, and around 50 or less were available for EMOINT.

5. Discussion

Our work has demonstrated the feasibility and potential of zero-shot classification in the context of building a fine-grained emotion detection dataset for literature analysis. Unlike previous research that focused on the direct use of zero-shot classification, we combined it with common semi-supervised learning approaches to construct a dataset for training supervised models. By leveraging the zero-shot NLI model, we ranked a large set of unsupervised sentences for each emotion and selected the most likely positive examples as weakly labeled data. These examples, along with their negative counterparts, were utilized to build binary pseudo-labeling models. By classifying all examples using these emotion-specific pseudo-labeling models, we obtained a multi-label dataset. To the best of our knowledge, this is the first publicly released dataset for multi-label fine-grained emotion detection in literature, with the most extensive emotion taxonomy of any dataset available, incorporating emotion labels not present in any other dataset. This is also the largest emotion detection dataset available and is two orders of magnitude bigger than the previously available literature dataset (Tales).

In the realm of emotion detection, our study showcases the possibility of having a fine-grained emotion detection dataset specifically tailored for literature. The supervised models trained on our dataset achieved promising results on our human-labeled gold dataset, even without substantial adaptation for handling noisy data and without hyperparameter tuning. The F1 score of 0.59 (Table 8) is similar to what is expected with crowdsourced multi-label emotion datasets (see Section 1.3 and Table 11). This is the case despite the fact that our dataset has more fine-grained labels and thus would be expected to be more challenging to model. It is an unexpected result that a semi-supervised dataset would provide results this good, especially even before adding any techniques for handling noise. This validates our semi-supervised approach.

The results comparing RoBERTa to BERT and the different model sizes (Table 10) suggest diminishing returns from increasing the quality of the underlying pretrained large language model, which can suggest better returns for focusing efforts on improving the dataset (e.g., filtering noisy examples) or the ability to learn from it (e.g., handling noisy labels).

Furthermore, our experiments in transfer learning, including zero-shot (Section 4.3) and few-shot scenarios (Section 4.4), reveal the transferability of the learned emotion detection capabilities. Models trained on our automatically generated dataset perform comparably to models trained on GoEmotions, a crowdsourced dataset. In the zero-shot scenario, for the literature dataset Tales, we achieve 93% of the Cross-Validation F1-score (Table 13). In the few-shot fine-tuning scenario, transfer models trained on our dataset surpass the models trained only on the Tales examples until the 200 examples threshold (Table 14). Both these transfer learning experiments further help validate the quality of our dataset and thus validate the approach used to create it.

Although the use of soft labels did not improve the results when evaluated on the gold subset of our dataset, all transfer experiments show superior results for the model trained

with soft labels, implying better transferability with soft labels. If the ultimate use of any model trained on our dataset is transfer learning, using the soft labels seems preferable.

6. Conclusions

In conclusion, this work proposed a novel semi-supervised approach for dataset creation, leveraging NLI zero-shot classification to construct a balanced multi-label fine-grained emotion detection dataset. We have demonstrated the effectiveness of our proposed approach, which compares favorably with crowdsourcing. We showed good performance for a model trained on our dataset when evaluated on a human labeled gold set. We also showed the transferability of emotion detection learned from our dataset to public benchmark datasets. We have provided a novel dataset and taxonomy for emotion detection with 38 emotion labels and made a model trained on our dataset public. We opened avenues for further exploration and improvement in emotion detection research, within the domain of literature analysis, cultural heritage analysis, and more broadly, any domain with a high cost of annotation.

The publication of our dataset and model will contribute to future research in emotion detection and in literature analysis. Within the scope of our project, Odeuropa (https://odeuropa.eu (accessed on 23 June 2023)), we will employ our model to investigate the relationship between olfactory sensations and emotions. Researchers can now conduct similar studies using our model, our dataset, or derived datasets. The size of our dataset allows for the creation of expert-annotated subsets that are still sufficiently large to train supervised fine-grained emotion detection models effectively.

From a methodological standpoint, there is ample room for improvement. Firstly, the emotion taxonomy can be expanded and refined to include additional categories such as "distrust" or "loneliness," which are significant for the study of the human condition. Secondly, data selection steps should be adjusted to address incomplete sentences and noisy sentences. More sophisticated token-counting approaches, such as using a tokenizer instead of simple whitespace delimitation, would be beneficial. Thirdly, incorporating techniques from related work that focus on improving zero-shot classification through ensembles of prompts and models could enhance the performance of our approach. Additionally, fine-tuning NLI with a few labeled examples (few-shot learning) has shown promise in related work and could be explored further, particularly if a small amount of human labeling is feasible, such as 5 to 10-shot per class. Fourthly, when constructing our initial binary models, more effort can be dedicated to reducing label noise or mitigating its impact beyond label smoothing. Minimal tuning of parameters with access to some labeled data could be explored. At this stage, incorporating hard negative example mining, which involves selecting negative examples with high semantic similarity to positive examples, could be beneficial. This approach forces the classifier to learn the expression of specific emotions and disregard surface regularities. Finally, while large-scale human annotation remains prohibitively expensive, recent advances in large language models could enable the utilization of chatbot assistants for performing the annotation task. Multiple assistants could simulate multiple annotators to some extent, offering a potential solution for this challenge. Immediate future work can also begin on handling whatever noise is present in our dataset. Noisy labels are a common issue with semi-supervised techniques for creating datasets, and many approaches have been presented over the years to handle it. These include both data filtering (removing likely noisy examples) and robust model training procedures (learning from noisy labels). Our dataset provides a good additional platform for the evaluation of these approaches.

Author Contributions: Conceptualization, L.R.; Data curation, L.R.; Funding acquisition, D.M.; Methodology, L.R.; Software, L.R.; Supervision, D.M.; Validation, L.R.; Writing—original draft, L.R.; Writing—review and editing, D.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the European Union through the Odeuropa EU H2020 project under grant agreement No 101004469.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset is available at https://zenodo.org/record/7883954. The models are available at https://huggingface.co/lrei/roberta-large-emolit, https://huggingface.co/lrei/roberta-base-emolit, and https://huggingface.co/lrei/distilroberta-base-emolit. The code to train the model is available at https://github.com/lrei/emolit_train.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Label Maps

Table A1 includes the output label maps from our taxonomy to the benchmark datasets used in Section 4.3. Labels not included are not used. See also Table 1 and Figure 4. Mapping to GoEmotions is one-to-one except for "remorse" which maps to "guilt" (simple label name change, the definition is the same) and that GoEmotions "realization" is mapped to "surprise".

Table A1. Label mappings from benchmark datasets to ours.

Tales	**EmoLit**
angry-disgusted	anger, annoyance, disapproval, disgust
happy	excitement, amusement, joy, relief, gratitude, optimism
fearful	fear, nervousness
sad	disappointment, despair, sadness, grief
surprised	surprise
ISEAR	**EmoLit**
fear	fear, nervousness
shame	embarrassment
guilt	guilt
disgust	disgust
anger	anger, annoyance, frustration
joy	approval, relief, gratitude, joy, optimism
sadness	grief, sadness
EMOINT	**EmoLit**
anger	anger, annoyance
joy	joy
fear	fear, nervousness
sadness	boredom, despair, sadness

Appendix B. NLI Hypothesis Comparison

Although plenty of results have been published on using NLI to perform zero-shot classification (see Section 1), since our purpose was to build a multi-label literature emotion, we conducted our own evaluation in a setting that is as close to ours as possible to give us an idea of how it would perform for the purpose of this work. This appendix shows the results of evaluating our zero-shot classifier on the Tales dataset. Since we are building a multi-label dataset, where each emotion is considered independent of all others, and Tales is a multiclass dataset, we consider all positive examples of a class in Tales to be negative examples of the other classes. Here we map "fearful" to "fear", "happy" to "joy", "sadness" to "sad". Table A2 shows these results with different thresholds for the probability of

the hypothesis being true given the premise. The trend is that the long hypothesis has a higher recall and the short hypothesis has a higher precision. The idea behind using the short hypothesis for ranking was to ensure that the top ranked examples were mostly true positives. Using the long hypothesis for re-ranking was a way to ensure diversity of the selected positive examples.

Table A2. NLI hypothesis binary evaluation at different thresholds on Tales: precision (P) recall (R), and F1-score per emotion.

Threshold	Hypothesis	Fear			Joy			Sadness		
		P	R	F1	P	R	F1	P	R	F1
0.5	Short	0.21	0.98	0.35	0.81	0.96	0.88	0.33	0.99	0.5
	Long	0.16	1.00	0.27	0.39	1.00	0.56	0.39	1.0	0.56
0.6	Short	0.22	0.97	0.36	0.83	0.94	0.88	0.35	0.99	0.51
	Long	0.16	1.00	0.28	0.39	1.00	0.56	0.32	1.00	0.49
0.7	Short	0.24	0.96	0.38	0.84	0.92	0.88	0.36	0.99	0.53
	Long	0.17	1.00	0.28	0.4	1.00	0.57	0.33	0.99	0.5
0.8	Short	0.27	0.93	0.42	0.86	0.87	0.87	0.39	0.98	0.55
	Long	0.17	1.00	0.29	0.42	1.00	0.59	0.35	0.99	0.52
0.9	Short	0.36	0.89	0.52	0.87	0.76	0.81	0.45	0.92	0.6
	Long	0.18	1.00	0.31	0.46	0.99	0.63	0.38	0.98	0.55
0.95	Short	0.51	0.75	0.6	0.88	0.64	0.74	0.53	0.88	0.66
	Long	0.19	1.0	0.32	0.53	0.99	0.69	0.42	0.95	0.58

Appendix C. Nostalgia

During exploration of the taxonomy, we investigated if it was possible to have a separate "nostalgia" and "homesickness" emotions. The dictionary definition of nostalgia is "a wistful or excessively sentimental yearning for return to or of some past period or irrecoverable condition", and often includes expressions of nostalgia towards things such as products, TV shows, and facilities. Nevertheless, the word was originally coined to describe acute homesickness and is still used in that context (https://www.merriam-webster.com/dictionary/nostalgia (accessed on 23 June 2023)). In our early explorations that lead to work, the two emotions were very easy to separate on social media, but when applied to Project Gutenberg Books, over 70% of examples weakly labeled as nostalgia were actual examples of homesickness and most of the others were strongly associated with locations. We decided to merge the two just after pseudo labeling in Section 2.5 by taking the max probability of either and assigning it to the label "nostalgia". This is why it is the biggest class in our dataset. Otherwise, that would more naturally be "disappointment".

References

1. Picard, R.W. *Affective Computing*; MIT Press: Cambridge, MA, USA, 2000.
2. Johansen, J.D. Feelings in literature. *Integr. Psychol. Behav. Sci.* **2010**, *44*, 185–196. [CrossRef] [PubMed]
3. Oatley, K. Emotions and the story worlds of fiction. In *Narrative Impact*; Psychology Press: Mahwah, NJ, USA, 2003; pp. 39–69.
4. Hogan, P.C. *What Literature Teaches Us about Emotion*; Cambridge University Press: Cambridge, UK, 2011.
5. Frevert, U. *Emotions in History—Lost and Found*; Central European University Press: Budapest, Hungary, 2011.
6. Massri, M.B.; Novalija, I.; Mladenić, D.; Brank, J.; Graça da Silva, S.; Marrouch, N.; Murteira, C.; Hürriyetoğlu, A.; Šircelj, B. Harvesting Context and Mining Emotions Related to Olfactory Cultural Heritage. *Multimodal Technol. Interact.* **2022**, *6*, 57. [CrossRef]
7. Alm, C.O.; Roth, D.; Sproat, R. Emotions from text: Machine learning for text-based emotion prediction. In Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Vancouver, BC, Canada, 6–8 October 2005; pp. 579–586.
8. Ekman, P. An argument for basic emotions. *Cogn. Emot.* **1992**, *6*, 169–200. [CrossRef]
9. Williams, L.; Arribas-Ayllon, M.; Artemiou, A.; Spasić, I. Comparing the utility of different classification schemes for emotive language analysis. *J. Classif.* **2019**, *36*, 619–648. [CrossRef]
10. Öhman, E. Emotion Annotation: Rethinking Emotion Categorization. In Proceedings of the DHN Post-Proceedings, Riga, Latvia, 21–23 October 2020; pp. 134–144.

11. Bostan, L.A.M.; Klinger, R. An Analysis of Annotated Corpora for Emotion Classification in Text. In Proceedings of the 27th International Conference on Computational Linguistics. Association for Computational Linguistics, Santa Fe, NM, USA, 20–26 August 2018; pp. 2104–2119.
12. Mohammad, S.; Bravo-Marquez, F.; Salameh, M.; Kiritchenko, S. SemEval-2018 Task 1: Affect in Tweets. In Proceedings of the 12th International Workshop on Semantic Evaluation, New Orleans, LA, USA, 5–6 June 2018; pp. 1–17. [CrossRef]
13. Demszky, D.; Movshovitz-Attias, D.; Ko, J.; Cowen, A.; Nemade, G.; Ravi, S. GoEmotions: A Dataset of Fine-Grained Emotions. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 4040–4054.
14. Kim, E.; Klinger, R. Who feels what and why? annotation of a literature corpus with semantic roles of emotions. In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 20–26 August 2018; pp. 1345–1359.
15. Menini, S.; Paccosi, T.; Tonelli, S.; Van Erp, M.; Leemans, I.; Lisena, P.; Troncy, R.; Tullett, W.; Hürriyetoğlu, A.; Dijkstra, G.; et al. A multilingual benchmark to capture olfactory situations over time. In Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change, Dublin, Ireland, 26–27 May 2022; pp. 1–10.
16. Rei, L.; Mladenic, D.; Dorozynski, M.; Rottensteiner, F.; Schleider, T.; Troncy, R.; Lozano, J.S.; Salvatella, M.G. Multimodal metadata assignment for cultural heritage artifacts. *Multimed. Syst.* **2023**, *29*, 847–869. [CrossRef]
17. Pita Costa, J.; Rei, L.; Stopar, L.; Fuart, F.; Grobelnik, M.; Mladenić, D.; Novalija, I.; Staines, A.; Pääkkönen, J.; Konttila, J.; et al. NewsMeSH: A new classifier designed to annotate health news with MeSH headings. *Artif. Intell. Med.* **2021**, *114*, 102053. [CrossRef] [PubMed]
18. Dagan, I.; Glickman, O.; Magnini, B. The pascal recognising textual entailment challenge. In Proceedings of the Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment: First PASCAL Machine Learning Challenges Workshop, MLCW 2005, Southampton, UK, 11–13 April 2005; pp. 177–190.
19. Yin, W.; Hay, J.; Roth, D. Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 3914–3923. [CrossRef]
20. Andreevskaia, A.; Bergler, S. CLaC and CLaC-NB: Knowledge-based and corpus-based approaches to sentiment tagging. In Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), Prague, Czech Republic, 23–24 June 2007; pp. 117–120.
21. Bermingham, A.; Smeaton, A.F. A study of inter-annotator agreement for opinion retrieval. In Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Boston, MA USA, 19–23 July 2009; pp. 784–785.
22. Russo, I.; Caselli, T.; Rubino, F.; Boldrini, E.; Martínez-Barco, P. EMOCause: An Easy-adaptable Approach to Extract Emotion Cause Contexts. In Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011), Portland, OR, USA, 24 June 2011; pp. 153–160.
23. Mohammad, S. A practical guide to sentiment annotation: Challenges and solutions. In Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, San Diego, CA, USA, 16 June 2016; pp. 174–179.
24. Plutchik, R. A general psychoevolutionary theory of emotion. In *Theories of Emotion*; Elsevier: Amsterdam, The Netherlands, 1980; pp. 3–33.
25. Scherer, K.R.; Wallbott, H.G. Evidence for universality and cultural variation of differential emotion response patterning. *J. Personal. Soc. Psychol.* **1994**, *66*, 310. [CrossRef] [PubMed]
26. Öhman, E.; Pàmies, M.; Kajava, K.; Tiedemann, J. XED: A Multilingual Dataset for Sentiment Analysis and Emotion Detection. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain, 8–13 December 2020; pp. 6542–6552. [CrossRef]
27. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.u.; Polosukhin, I. Attention is All you Need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
28. 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, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 7 June 2019; pp. 4171–4186. [CrossRef]
29. Kocoń, J.; Cichecki, I.; Kaszyca, O.; Kochanek, M.; Szydło, D.; Baran, J.; Bielaniewicz, J.; Gruza, M.; Janz, A.; Kanclerz, K.; et al. ChatGPT: Jack of all trades, master of none. *Inf. Fusion* **2023**, *99*, 101861. [CrossRef]
30. Ameer, I.; Bölücü, N.; Siddiqui, M.H.F.; Can, B.; Sidorov, G.; Gelbukh, A. Multi-label emotion classification in texts using transfer learning. *Expert Syst. Appl.* **2023**, *213*, 118534. [CrossRef]
31. Alhuzali, H.; Ananiadou, S. SpanEmo: Casting Multi-label Emotion Classification as Span-prediction. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Online, 19–23 April 2021; pp. 1573–1584. [CrossRef]

32. Basile, A.; Pérez-Torró, G.; Franco-Salvador, M. Probabilistic Ensembles of Zero- and Few-Shot Learning Models for Emotion Classification. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), Online, 1–3 September 2021; pp. 128–137.
33. Plaza-del Arco, F.M.; Martín-Valdivia, M.T.; Klinger, R. Natural Language Inference Prompts for Zero-shot Emotion Classification in Text across Corpora. In Proceedings of the 29th International Conference on Computational Linguistics, Gyeongju, Republic of Korea, 12–17 October 2022; pp. 6805–6817.
34. Tesfagergish, S.G.; Kapočiūtė-Dzikienė, J.; Damaševičius, R. Zero-Shot Emotion Detection for Semi-Supervised Sentiment Analysis Using Sentence Transformers and Ensemble Learning. *Appl. Sci.* **2022**, *12*, 8662. [CrossRef]
35. Gera, A.; Halfon, A.; Shnarch, E.; Perlitz, Y.; Ein-Dor, L.; Slonim, N. Zero-Shot Text Classification with Self-Training. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, United Arab Emirates, 7–11 December 2022; pp. 1107–1119.
36. Peterson, J.C.; Battleday, R.M.; Griffiths, T.L.; Russakovsky, O. Human uncertainty makes classification more robust. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9617–9626.
37. Van Engelen, J.E.; Hoos, H.H. A survey on semi-supervised learning. *Mach. Learn.* **2020**, *109*, 373–440. [CrossRef]
38. El Gayar, N.; Schwenker, F.; Palm, G. A study of the robustness of KNN classifiers trained using soft labels. In Proceedings of the Artificial Neural Networks in Pattern Recognition: Second IAPR Workshop, ANNPR 2006, Ulm, Germany, 31 August–2 September 2006; pp. 67–80.
39. Thiel, C. Classification on soft labels is robust against label noise. In Proceedings of the Knowledge-Based Intelligent Information and Engineering Systems: 12th International Conference, KES 2008, Zagreb, Croatia, 3–5 September 2008; pp. 65–73.
40. Galstyan, A.; Cohen, P.R. Empirical comparison of "hard" and "soft" label propagation for relational classification. In Proceedings of the International Conference on Inductive Logic Programming, Corvallis, OR, USA, 25–27 October 2007; pp. 98–111.
41. Zhao, Z.; Wu, S.; Yang, M.; Chen, K.; Zhao, T. Robust machine reading comprehension by learning soft labels. In Proceedings of the 28th International Conference on Computational Linguistics, Online, 8–13 December 2020; pp. 2754–2759.
42. Bird, S.; Klein, E.; Loper, E. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2009.
43. Kiss, T.; Strunk, J. Unsupervised multilingual sentence boundary detection. *Comput. Linguist.* **2006**, *32*, 485–525. [CrossRef]
44. Joulin, A.; Grave, E.; Bojanowski, P.; Mikolov, T. Bag of Tricks for Efficient Text Classification. *arXiv* **2016**, arXiv:1607.01759.
45. Joulin, A.; Grave, E.; Bojanowski, P.; Douze, M.; Jégou, H.; Mikolov, T. FastText.zip: Compressing text classification models. *arXiv* **2016**, arXiv:1612.03651.
46. Broder, A.Z. On the resemblance and containment of documents. In Proceedings of the Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171), Positano, Italy, 11–13 June 1997; pp. 21–29.
47. Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; Zettlemoyer, L. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 7871–7880. [CrossRef]
48. Williams, A.; Nangia, N.; Bowman, S. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Seattle, WA, USA, 10–15 July 2022; pp. 1112–1122.
49. Wang, W.; Wei, F.; Dong, L.; Bao, H.; Yang, N.; Zhou, M. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 5776–5788.
50. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Hong Kong, China, 3–7 November 2019.
51. Porter, M.F. An algorithm for suffix stripping. *Program* **1980**, *14*, 130–137. [CrossRef]
52. Lee, D.H. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In Proceedings of the Workshop on Challenges in Representation Learning, ICML, Atlanta, GA, USA, 16–21 June 2013; Volume 3, p. 896.
53. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826. [CrossRef]
54. Sechidis, K.; Tsoumakas, G.; Vlahavas, I. On the Stratification of Multi-label Data. In Proceedings of the Machine Learning and Knowledge Discovery in Databases, Athens, Greece, 5–9 September 2011; Gunopulos, D., Hofmann, T., Malerba, D., Vazirgiannis, M., Eds.; Springer: Berlin/Heidelberg, Germany, 2011; pp. 145–158.
55. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
56. Sennrich, R.; Haddow, B.; Birch, A. Neural Machine Translation of Rare Words with Subword Units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, 7–12 August 2016; pp. 1715–1725. [CrossRef]
57. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. *Language Models Are Unsupervised Multitask Learners*; Technical Report; OpenAI: San Francisco, CA, USA, 2019.

58. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
59. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv* **2019**, arXiv:1910.01108.
60. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
61. Schutze, H.; Manning, C.D.; Raghavan, P. *Introduction to Information Retrieval*; Cambridge University Press: Cambridge, UK, 2008; p. 281.
62. Mohammad, S.; Bravo-Marquez, F. Emotion Intensities in Tweets. In Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017), Vancouver, BC, Canada, 3–4 August 2017; pp. 65–77. [CrossRef]

Chapter 5

Conclusions

In our work, we tackled the significant challenge of applying machine learning in knowledge-intensive resource-scarce situations where a very limited number of labeled examples are available. Using Cultural Heritage as the application domain, we first showed how to use transfer learning to overcome the challenge. We also applied the same principles to the healthcare domain by learning how to classify news articles from scientific articles. Our efforts led to the development of a sophisticated multimodal multilingual multitask approach for predicting missing metadata in digitized cultural heritage archives that effectively predicts missing metadata within digitized cultural heritage archives. We then expanded our work beyond transfer learning and into semi-supervised methods that allowed us to build a model and dataset for a highly complex task, emotion detection in literature, without requiring any labeled examples, negating one of the biggest challenges in this domain.

We first showed that it was possible to infer missing metadata from text descriptions of digitized cultural artifacts. Using data retrieved from a single archive containing thousands of text descriptions of digitized silk fabrics, we showed that we could learn a model to infer metadata for different expert-selected important categories: "Production Technique", "Material Used", "Production Place", and "Production Date". We achieved an average accuracy of 94%. Having leveraged a thesaurus to map metadata across archives, we then showed that a model that learned exclusively from the data in this first archive could be used to infer metadata of objects in a different archive in the same language with 60% average accuracy. We showed that the same model could be used to infer metadata for objects in an archive with text descriptions written in a different language, with an average accuracy of 55%. Having shown that transfer across archives was possible, we combined data from multiple archives, multiple languages, and multiple tasks in a single dataset. We learned a multilingual multi-task model using simple mixed training and obtained an average accuracy of 89% and an average macro F1 of 86% on a held-out test set of digitized objects from multiple archives in 4 different languages. Additionally, taking advantage of a different thesaurus, the Medical Subject Headings (MeSH), we showed that we could learn from the metadata of scientific articles in online archives to categorize news articles with a 56% macro-averaged F1 score.

The next stage in our work on inferring metadata for digitized cultural heritage objects was to go beyond using just text descriptions of the objects. We trained tabular classifiers that inferred the value of each category of metadata using the values of the other categories. This approach achieved an average macro F1 of 56% on the held-out test set previously mentioned. Although this number is significantly worse than the text classifier, it considers all objects, while the text classifier only considers objects with text descriptions. These are only available for 39% of the digitized silk fabrics in our dataset. The same phenomenon occurs when considering images, although an image classifier only achieves an average F1

score of 58%, almost 97% of digitized objects have at least one image associated with them. We used this knowledge to develop a multimodal approach based on late fusion. When considering all objects in our dataset, the multimodal classifier achieves an average F1 score of 74% which is significantly higher than the 56% average F1 of the tabular classifier, the 52% of the image classifier, or the 49% of the text classifier. We showed that each modality contributes to the multimodal classifier and that using all three modalities is better than using any combination of two modalities.

In the final part of this work, we considered cultural heritage in the form of literature and the problem of text classification without any human-labeled data. Using multi-label fine-grained emotion detection as the task to accomplish, we created a dataset with 38 emotion labels automatically using semi-supervised methods. After training a classifier on this dataset, we evaluated the results on a manually labeled test set. The result was a macro F1 score of 59%, a result in line with, or better than, results shown for human-labeled datasets for multi-label emotion detection. We showed that our classifier learned a representation of emotion that was transferable to other datasets and domains. Lastly, we presented an analysis of emotion correlation and sentiment-emotion correlation within the large dataset we made publicly available. This analysis should help inform future work, especially in literature, as we showed that a significant number of sentences are written to have simultaneously positive and negative emotions.

The consistent theme across these studies is the innovative use of machine learning techniques to enhance unstructured data usability in expert domains, especially cultural heritage, through a common methodology involving the use of controlled vocabularies, representation learning, and transfer learning. The applied nature of this research has clear implications for fields such as digital humanities, public health, and cultural heritage management. By automating the classification of data, these studies contribute to more efficient knowledge discovery processes, which are crucial in managing large-scale databases and enhancing information retrieval systems. The methodology developed is not confined to cultural heritage but serves as a template for other knowledge-intensive, data-scarce fields, including those requiring handling multimodal data. Before our work, the option was not to have metadata for certain artifacts, metadata for news articles, or fine-grained emotions associated with book passages. This meant that the platforms, such as observatories and exploratory search engines relied exclusively on keyword search to find and organize relevant items relevant items or to provide analysis of certain topics or along certain dimensions. After our work, these platforms could rely on mostly accurate metadata. Beyond the quantitative analysis we presented, formal and informal qualitative evaluations were carried out in the context of the larger research projects of which our work was a part. Partners, reviewers, and audiences generally agreed on the usefulness of the outcomes enabled in part by our work in predicting metadata. This metadata was, however, not perfectly accurate. For instance, in the SILKNOW Deliverable *D7.2. Testing report in a real scenario*, expert users performing metadata-based queries pointed out that some results were inaccurate. At the same time, some expert users couldn't evaluate other results e.g. the classifier provided a location for a silk object but the text itself did not mention any location. The latter is intended behavior, if nothing else, a classifier can learn conditional label priors based on data regularities such as "if the text is written in Spanish and no other location is mentioned, predict production location to be Spain".

5.1 Limitations

While the developed methodology has significantly advanced the prediction of missing metadata, enhancing applications in the process, its effectiveness is contingent upon the

quality of the input data. We acknowledge that label errors are present in all utilized training datasets; however, an in-depth estimation and analysis of these errors and their impact on the transfer results were not conducted. Additionally, the success of transferring metadata across different archives not only depends on the data quality but also on the similarity between the source and target archives—a factor that was not extensively investigated. Similarly, our efforts to address underrepresented classes in the datasets were not applied universally or explored comprehensively.

The methodology heavily relies on controlled vocabularies for training, which, while improving consistency and facilitating data integration across different organizations, also introduces a dependency on the completeness and accuracy of these vocabularies. In certain domains, appropriate controlled vocabularies may not exist. For silk fabrics, the thesaurus was not pre-existing and had to be developed concurrently with our research.

Moreover, our methodology focuses on classification tasks. While it offers prospects for adaptation to data extraction tasks, it does not encompass regression and we did not deal with time series data. Both of these are critical in expert fields predominantly characterized by numeric data, including manufacturing, energy consumption monitoring, predictive maintenance, and environmental monitoring.

5.2 Future Work

Having handled underrepresented classes in the course of our work on digitized silk fabrics, an avenue we find interesting for continuing our work in healthcare is to investigate and potentially improve its performance in detecting rare conditions, such as Prader-Willi or Rett syndrome, using a similar approach. The first step in this direction, classifying rare diseases in scientific abstracts and news articles was published as *Automatic Classification and Visualization of Text Data on Rare Diseases* (Rei et al., 2024) in the Journal of Personalized Medicine, Volume 14, Issue 5, in May 2024, and belongs to the Special Issue Artificial Intelligence and Data Integration in Precision Health. The journal has an impact factor of 3.4 and is indexed in the Science Citation Index Expanded. With rare diseases being, by definition, rare, we see the potential for the exploration and adoption of methods for class balancing such as the Distribution-Balanced Loss (Huang et al., 2021; T. Wu et al., 2020).

In our research on digitized silk fabrics, we have identified noisy labels as a challenge. Similarly, our work on emotion detection relies on a Semi-supervised learning approach that is susceptible to noise accumulation or confirmation bias, where initial inaccuracies in model predictions on unlabeled data can be propagated to the final model during training (Arazo et al., 2020; Tarvainen & Valpola, 2017; Yarowsky, 1995; Z. Zhang et al., 2016). Currently, the leading approaches in noisy label learning predominantly utilize semi-supervised learning techniques (W. Chen et al., 2023; Li et al., 2020; Xiao et al., 2023). Semi-supervised learning also plays a foundational role in unsupervised domain adaptation (Ramponi & Plank, 2020), which is necessary in knowledge-intensive domains since it is costly and difficult to get domain experts to label data in the target domain beyond a small test set as is the case in our healthcare study. Thus, enhancing our methodology by more deeply integrating SSL—both to improve handling of noisy labels and to refine our domain adaptation capabilities—represents a promising and synergistic direction for future research efforts.

For what concerns multimodal metadata assignment in cultural heritage, it would be interesting to try an approach similar to CLIP (Radford et al., 2021) and CLIP-Art (Conde & Turgutlu, 2021) to create the representation for image classification using supervision from natural language before multimodal classification. CLIP (Contrastive Language-Image Pre-training) consists of a model trained image-text pairs, learning to understand

and categorize images based on textual descriptions. CLIP uses a contrastive learning framework that enhances its ability to correlate text embeddings with visual features, enabling robust zero-shot and few-shot learning capabilities. This allows CLIP to classify images into unseen categories based solely on textual descriptions, making it highly effective for applications where specific training data is scarce or unavailable as is the case in CH. The recent advances in Multimodal Large Language Models (J. Wu et al., 2023) also make fine-tuning these models to the CH domain a promising avenue for future multimodal work.

The methodology we proposed addresses problems necessitating expert knowledge where this expertise is synthesized into class labels. In our parallel study on water-related extreme events (Pita Costa et al., 2024), we dealt with expert knowledge represented as numeric time series. We envision significant potential in extending the methods presented in this thesis from classification to regression, which could have a substantial impact in numerically intensive fields such as Engineering, Economics, and Finance.

References

Alba, E., Gaitán, M., León, A., Mladenić, D., & Brank, J. (2022). Weaving words for textile museums: The development of the linked silknow thesaurus. *Heritage Science*, *10*(1), 1–14.

Alm, C. O., Roth, D., & Sproat, R. (2005). Emotions from text: Machine learning for text-based emotion prediction. *HLT/EMNLP 2005, Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 6-8 October 2005, Vancouver, British Columbia, Canada*, 579–586.

Arazo, E., Ortego, D., Albert, P., O'Connor, N. E., & McGuinness, K. (2020). Pseudo-labeling and confirmation bias in deep semi-supervised learning. *2020 International Joint Conference on Neural Networks, IJCNN 2020, Glasgow, United Kingdom, July 19-24, 2020*, 1–8. https://doi.org/10.1109/IJCNN48605.2020.9207304

Baltrusaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. *41*(2), 423–443. https://doi.org/10.1109/TPAMI.2018.2798607

Belhi, A., Bouras, A., & Foufou, S. (2018). Leveraging known data for missing label prediction in cultural heritage context. *Applied Sciences*, *8*(10). https://doi.org/10.3390/app8101768

Bengio, Y. (2012). Deep learning of representations for unsupervised and transfer learning. In I. Guyon, G. Dror, V. Lemaire, G. W. Taylor, & D. L. Silver (Eds.), *Unsupervised and transfer learning - workshop held at ICML 2011, bellevue, washington, usa, july 2, 2011* (pp. 17–36). JMLR.org. http://proceedings.mlr.press/v27/bengio12a.html

Caruana, R. (1997). Multitask learning. *Machine Learning*, *28*(1), 41–75. https://doi.org/10.1023/A:1007379606734

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In B. Krishnapuram, M. Shah, A. J. Smola, C. C. Aggarwal, D. Shen, & R. Rastogi (Eds.), *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, san francisco, ca, usa, august 13-17, 2016* (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

Chen, W., Zhu, C., & Li, M. (2023). Sample prior guided robust model learning to suppress noisy labels. In D. Koutra, C. Plant, M. G. Rodriguez, E. Baralis, & F. Bonchi (Eds.), *Machine learning and knowledge discovery in databases: Research track - european conference, ECML PKDD 2023, turin, italy, september 18-22, 2023, proceedings, part II* (pp. 3–19). Springer. https://doi.org/10.1007/978-3-031-43415-0_1

Cheng, Y.-Y., & Xia, Y. (2023). A systematic review of methods for aligning, mapping, merging taxonomies in information sciences. *Journal of Documentation*, *ahead-of-print*(ahead-of-print). https://doi.org/10.1108/JD-01-2023-0003

Conde, M. V., & Turgutlu, K. (2021). Clip-art: Contrastive pre-training for fine-grained art classification. *IEEE Conference on Computer Vision and Pattern Recognition*

Workshops, CVPR Workshops 2021, virtual, June 19-25, 2021, 3956–3960. https://doi.org/10.1109/CVPRW53098.2021.00444

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In D. Jurafsky, J. Chai, N. Schluter, & J. R. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics, ACL 2020, online, july 5-10, 2020* (pp. 8440–8451). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.747

Daumé III, H. (2007). Frustratingly easy domain adaptation. In J. Carroll, A. van den Bosch, & A. Zaenen (Eds.), *ACL 2007, proceedings of the 45th annual meeting of the association for computational linguistics, june 23-30, 2007, prague, czech republic* (pp. 256–263). The Association for Computational Linguistics. https://aclanthology.org/P07-1033/

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A. S., Nemade, G., & Ravi, S. (2020). Goemotions: A dataset of fine-grained emotions. In D. Jurafsky, J. Chai, N. Schluter, & J. R. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics, ACL 2020, online, july 5-10, 2020* (pp. 4040–4054). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.372

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies, NAACL-HLT 2019, minneapolis, mn, usa, june 2-7, 2019, volume 1 (long and short papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/n19-1423

Dorozynski, M., Clermont, D., & Rottensteiner, F. (2019). Multi-task deep learning with incomplete training samples for the image-based prediction of variables describing silk fabrics. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *IV-2/W6*, 47–54. https://doi.org/10.5194/isprs-annals-IV-2-W6-47-2019

Ehrhart, T., Lisena, P., & Troncy, R. (2021). KG explorer: A customisable exploration tool for knowledge graphs. In P. Lambrix, C. Pesquita, & V. Wiens (Eds.), *Proceedings of the sixth international workshop on the visualization and interaction for ontologies and linked data co-located with the 20th international semantic web conference (ISWC 2021), virtual conference, 2021* (pp. 63–75). CEUR-WS.org. https://ceur-ws.org/Vol-3023/paper13.pdf

Ericsson, L., Gouk, H., Loy, C. C., & Hospedales, T. M. (2022). Self-supervised representation learning: Introduction, advances, and challenges. *IEEE Signal Processing Magazine*, *39*(3), 42–62. https://doi.org/10.1109/MSP.2021.3134634

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning* [http://www.deeplearningbook.org]. MIT Press.

Gottschall, J. (2012). *The storytelling animal: How stories make us human.* Houghton Mifflin Harcourt.

GRAHAM, J., & HAIDT, J. (2012). Sacred values and evil adversaries: A moral foundations approach. In *The social psychology of morality: Exploring the causes of good and evil* (pp. 11–32). American Psychological Association. https://doi.org/https://doi.org/10.1037/13091-001

Halevy, A. Y., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE Intell. Syst.*, *24*(2), 8–12. https://doi.org/10.1109/MIS.2009.36

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Identity mappings in deep residual networks. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *Computer vision - ECCV 2016 - 14th european conference, amsterdam, the netherlands, october 11-14, 2016, proceedings, part IV* (pp. 630–645). Springer. https://doi.org/10.1007/978-3-319-46493-0_38

Henderson, L. (2009). *Automated text classification in the dmoz hierarchy* (tech. rep.).

Huang, Y., Giledereli, B., Köksal, A., Özgür, A., & Ozkirimli, E. (2021). Balancing methods for multi-label text classification with long-tailed class distribution. In M.-F. Moens, X. Huang, L. Specia, & S. W.-t. Yih (Eds.), *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 8153–8161). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.643

Jafari, A. R., Heidary, B., Farahbakhsh, R., Salehi, M., & Jalili, M. (2021). Transfer learning for multi-lingual tasks - a survey. *CoRR*, *abs/2110.02052*. https://arxiv.org/abs/2110.02052

Johansen, J. D. (2010). Feelings in literature. *Integrative Psychological and Behavioral Science*, *44*(3), 185–196. https://doi.org/10.1007/s12124-009-9112-0

Joulin, A., Bojanowski, P., Mikolov, T., Jégou, H., & Grave, E. (2018). Loss in translation: Learning bilingual word mapping with a retrieval criterion. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Eds.), *Proceedings of the 2018 conference on empirical methods in natural language processing, brussels, belgium, october 31 - november 4, 2018* (pp. 2979–2984). Association for Computational Linguistics. https://doi.org/10.18653/v1/d18-1330

Kim, Y. (2014). Convolutional neural networks for sentence classification. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 conference on empirical methods in natural language processing, EMNLP 2014, october 25-29, 2014, doha, qatar, A meeting of sigdat, a special interest group of the ACL* (pp. 1746–1751). ACL. https://doi.org/10.3115/v1/d14-1181

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In P. L. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 25: 26th annual conference on neural information processing systems 2012. proceedings of a meeting held december 3-6, 2012, lake tahoe, nevada, united states* (pp. 1106–1114). https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

Léon, A., Gaitán, M., Sebastián, J., Pagán, E. A., & Insa, I. (2019). Silknow. designing a thesaurus about historical silk for small and medium-sized textile museums. In *Science and digital technology for cultural heritage-interdisciplinary approach to diagnosis, vulnerability, risk assessment and graphic information models* (pp. 187–190). CRC Press.

Li, J., Socher, R., & Hoi, S. C. H. (2020). Dividemix: Learning with noisy labels as semi-supervised learning. *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. https://openreview.net/forum?id=HJgExaVtwr

Liu, X., He, P., Chen, W., & Gao, J. (2019). Multi-task deep neural networks for natural language understanding. In A. Korhonen, D. R. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th conference of the association for computational linguistics, ACL 2019, florence, italy, july 28- august 2, 2019, volume 1: Long papers* (pp. 4487–4496). Association for Computational Linguistics. https://doi.org/10.18653/v1/p19-1441

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *CoRR*, *abs/1907.11692*. http://arxiv.org/abs/1907.11692

Massri, M. B., Novalija, I., Mladenić, D., Brank, J., Graça da Silva, S., Marrouch, N., Murteira, C., Hürriyetoğlu, A., & Šircelj, B. (2022). Harvesting context and mining emotions related to olfactory cultural heritage. *Multimodal Technologies and Interaction*, *6*(7). https://doi.org/10.3390/mti6070057

Midas: Meaningful integration of data, analytics and services [2016–2020]. (2016). https://doi.org/10.3030/727721

Mladenić, D. (1998). Turning yahoo into automatic web-page classifier. *13th European Conference on Artificial Intelligence*.

Mohammad, S. M., & Bravo-Marquez, F. (2017). Emotion intensities in tweets. In N. Ide, A. Herbelot, & L. Màrquez (Eds.), *Proceedings of the 6th joint conference on lexical and computational semantics, *sem @acm 2017, vancouver, canada, august 3-4, 2017* (pp. 65–77). Association for Computational Linguistics. https://doi.org/10.18653/v1/S17-1007

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). *Foundations of machine learning* (2nd ed.). MIT Press.

Oatley, K. (2003). Emotions and the story worlds of fiction. In *Narrative impact* (pp. 39–69). Psychology Press.

Odeuropa: Negotiating olfactory and sensory experiences in cultural heritage practice and research [2021–2023]. (2021). https://doi.org/10.3030/101004469

Oquab, M., Bottou, L., Laptev, I., & Sivic, J. (2014). Learning and transferring mid-level image representations using convolutional neural networks. *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, 1717–1724. https://doi.org/10.1109/CVPR.2014.222

Palagi, É. (2018). *Evaluating exploratory search engines: Designing a set of user-centered methods based on a modeling of the exploratory search process. (evaluation des moteurs de recherche exploratoire: Élaboration d'un corps de méthodes centrées utilisateurs, basées sur une modélisation du processus de recherche exploratoire)* (Doctoral dissertation). Université Côte d'Azur, France. https://tel.archives-ouvertes.fr/tel-01976017

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.*, *22*(10), 1345–1359. https://doi.org/10.1109/TKDE.2009.191

Pikuliak, M., Šimko, M., & Bieliková, M. (2021). Cross-lingual learning for text processing: A survey. *Expert Systems with Applications*, *165*, 113765. https://doi.org/https://doi.org/10.1016/j.eswa.2020.113765

Pita Costa, J., Rei, L., Bezak, N., Mikoš, M., Massri, M. B., Novalija, I., & Leban, G. (2024). Towards improved knowledge about water-related extremes based on news media information captured using artificial intelligence. *International Journal of Disaster Risk Reduction*, *100*, 104172. https://doi.org/https://doi.org/10.1016/j.ijdrr.2023.104172

Pita Costa, J., Rei, L., Stopar, L., Fuart, F., Grobelnik, M., Mladenić, D., Novalija, I., Staines, A., Pääkkönen, J., Konttila, J., Bidaurrazaga, J., Belar, O., Henderson, C., Epelde, G., Gabaráin, M. A., Carlin, P., & Wallace, J. (2021). Newsmesh: A new classifier designed to annotate health news with mesh headings. *Artificial Intelligence in Medicine*, *114*, 102053. https://doi.org/https://doi.org/10.1016/j.artmed.2021.102053

Plested, J., & Gedeon, T. (2022). Deep transfer learning for image classification: A survey. *CoRR*, *abs/2205.09904*. https://doi.org/10.48550/arXiv.2205.09904

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In M. Meila & T. Zhang (Eds.), *Proceedings of the 38th international conference on machine learning, ICML 2021, 18-24 july 2021, virtual event* (pp. 8748–8763). PMLR. http://proceedings.mlr.press/v139/radford21a.html

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training* (tech. rep.). OpenAI.

Ramachandram, D., & Taylor, G. W. (2017). Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, *34*(6), 96–108. https://doi.org/10.1109/MSP.2017.2738401

Ramponi, A., & Plank, B. (2020). Neural unsupervised domain adaptation in NLP - A survey. In D. Scott, N. Bel, & C. Zong (Eds.), *Proceedings of the 28th international conference on computational linguistics, COLING 2020, barcelona, spain (online), december 8-13, 2020* (pp. 6838–6855). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.603

Rankin, D., Black, M., Wallace, J., Mulvenna, M., Bond, R., & Cleland, B. (2017). The midas platform: Facilitating the utilisation of healthcare big data in northern ireland and beyond. *8th Annual Translational Medicine Conference*.

Rei, L., Mladenic, D., Dorozynski, M., Rottensteiner, F., Schleider, T., Troncy, R., Lozano, J. S., & Salvatella, M. G. (2023). Multimodal metadata assignment for cultural heritage artifacts. *Multimedia Systems*, *29*(2), 847–869. https://doi.org/10.1007/s00530-022-01025-2

Rei, L., & Mladenić, D. (2020). Cross-lingual text classification for inferring variables describing silk fabrics. *Weaving Europe Silk Heritage and digital technologies*, 177–190.

Rei, L., & Mladenić, D. (2023). Detecting fine-grained emotions in literature. *Applied Sciences*, *13*(13). https://doi.org/10.3390/app13137502

Rei, L., Pita Costa, J., & Zdolšek Draksler, T. (2024). Automatic classification and visualization of text data on rare diseases. *Journal of Personalized Medicine*, *14*(5). https://doi.org/10.3390/jpm14050545

Ruder, S. (2017). An overview of multi-task learning in deep neural networks. *CoRR*, *abs/1706.05098*. http://arxiv.org/abs/1706.05098

Ruder, S. (2019). *Neural transfer learning for natural language processing* (Doctoral dissertation). NUI Galway.

Ruotsalo, T., Aroyo, L., & Schreiber, G. (2009a). Knowledge-based linguistic annotation of digital cultural heritage collections. *IEEE Intell. Syst.*, *24*(2), 64–75. https://doi.org/10.1109/MIS.2009.32

Ruotsalo, T., Aroyo, L., & Schreiber, G. (2009b). Knowledge-based linguistic annotation of digital cultural heritage collections. *IEEE Intell. Syst.*, *24*(2), 64–75. https://doi.org/10.1109/MIS.2009.32

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. S., Berg, A. C., & Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, *115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Schleider, T., Troncy, R., Gaitan, M., Alba, E., & et al. (2021). The silknow knowledge graph. *Semantic Web Journal, Special Issue on Cultural Heritage and Semantic Web, March 2021, IOS Press*.

Sebastián Lozano, J., Alba Pagán, E., Martínez Roig, E., Gaitán Salvatella, M., León Muñoz, A., Sevilla Peris, J., Vernus, P., Puren, M., Rei, L., & Mladenič, D. (2023).

Open access to data about silk heritage: A case study in digital information sustainability. *Sustainability*, *15*(19). https://doi.org/10.3390/su151914340

Silknow. silk heritage in the knowledge society: From punched cards to big data, deep learning and visual / tangible simulations [2018–2021]. (2018). https://doi.org/10.3030/769504

Stefanini, M., Cornia, M., Baraldi, L., Corsini, M., & Cucchiara, R. (2019). Artpedia: A new visual-semantic dataset with visual and contextual sentences in the artistic domain. In E. Ricci, S. R. Bulò, C. Snoek, O. Lanz, S. Messelodi, & N. Sebe (Eds.), *Image analysis and processing - ICIAP 2019 - 20th international conference, trento, italy, september 9-13, 2019, proceedings, part II* (pp. 729–740). Springer. https://doi.org/10.1007/978-3-030-30645-8%5C_66

Tarvainen, A., & Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems*. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/68053af2923e00204c3ca7c6a3150cf7-Paper.pdf

van Erp, M., Tullett, W., Christlein, V., Ehrhart, T., Hürriyetoğlu, A., Leemans, I., Lisena, P., Menini, S., Schwabe, D., Tonelli, S., Troncy, R., & Zinnen, M. (2023). More than the Name of the Rose: How to Make Computers Read, See, and Organize Smells. *The American Historical Review*, *128*(1), 335–369. https://doi.org/10.1093/ahr/rhad141

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In T. Linzen, G. Chrupala, & A. Alishahi (Eds.), *Proceedings of the workshop: Analyzing and interpreting neural networks for nlp, blackboxnlp@emnlp 2018, brussels, belgium, november 1, 2018* (pp. 353–355). Association for Computational Linguistics. https://doi.org/10.18653/v1/w18-5446

Wang, Y., Yao, Q., Kwok, J. T., & Ni, L. M. (2020). Generalizing from a few examples: A survey on few-shot learning. *ACM Comput. Surv.*, *53*(3). https://doi.org/10.1145/3386252

Wilson, G., & Cook, D. J. (2020). A survey of unsupervised deep domain adaptation. *ACM Trans. Intell. Syst. Technol.*, *11*(5). https://doi.org/10.1145/3400066

Wu, J., Gan, W., Chen, Z., Wan, S., & Yu, P. S. (2023). Multimodal large language models: A survey. *2023 IEEE International Conference on Big Data (BigData)*, 2247–2256. https://doi.org/10.1109/BigData59044.2023.10386743

Wu, T., Huang, Q., Liu, Z., Wang, Y., & Lin, D. (2020). Distribution-balanced loss for multi-label classification in long-tailed datasets. In A. Vedaldi, H. Bischof, T. Brox, & J. Frahm (Eds.), *Computer vision - ECCV 2020 - 16th european conference, glasgow, uk, august 23-28, 2020, proceedings, part IV* (pp. 162–178). Springer. https://doi.org/10.1007/978-3-030-58548-8_10

Xiao, R., Dong, Y., Wang, H., Feng, L., Wu, R., Chen, G., & Zhao, J. (2023). Promix: Combating label noise via maximizing clean sample utility. *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*, 4442–4450. https://doi.org/10.24963/IJCAI.2023/494

Yang, Z., Dai, Z., Yang, Y., Carbonell, J. G., Salakhutdinov, R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32: Annual conference*

on neural information processing systems 2019, neurips 2019, december 8-14, 2019, vancouver, bc, canada (pp. 5754–5764). https://proceedings.neurips.cc/paper/2019/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. *33rd Annual Meeting of the Association for Computational Linguistics*, 189–196. https://doi.org/10.3115/981658.981684

Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. (2021). Dive into deep learning. *arXiv preprint arXiv:2106.11342*.

Zhang, C., Kaeser-Chen, C., Vesom, G., Choi, J., Kessler, M., & Belongie, S. J. (2019). The imet collection 2019 challenge dataset. *CoRR*, *abs/1906.00901*. http://arxiv.org/abs/1906.00901

Zhang, Y., & Yang, Q. (2017). An overview of multi-task learning. *National Science Review*, *5*(1), 30–43. https://doi.org/10.1093/nsr/nwx105

Zhang, Z., Ringeval, F., Dong, B., Coutinho, E., Marchi, E., & Schüller, B. (2016). Enhanced semi-supervised learning for multimodal emotion recognition. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5185–5189. https://doi.org/10.1109/ICASSP.2016.7472666

Bibliography

Publications Related to the Thesis

Journal Articles

Pita Costa, J., Rei, L., Bezak, N., Mikoš, M., Massri, M. B., Novalija, I., & Leban, G. (2024). Towards improved knowledge about water-related extremes based on news media information captured using artificial intelligence. *International Journal of Disaster Risk Reduction*, *100*, 104172. https://doi.org/https://doi.org/10.1016/j.ijdrr.2023.104172

Pita Costa, J., Rei, L., Stopar, L., Fuart, F., Grobelnik, M., Mladenić, D., Novalija, I., Staines, A., Pääkkönen, J., Konttila, J., Bidaurrazaga, J., Belar, O., Henderson, C., Epelde, G., Gabaráin, M. A., Carlin, P., & Wallace, J. (2021). Newsmesh: A new classifier designed to annotate health news with mesh headings. *Artificial Intelligence in Medicine*, *114*, 102053. https://doi.org/https://doi.org/10.1016/j.artmed.2021.102053

Rei, L., Mladenic, D., Dorozynski, M., Rottensteiner, F., Schleider, T., Troncy, R., Lozano, J. S., & Salvatella, M. G. (2023). Multimodal metadata assignment for cultural heritage artifacts. *Multimedia Systems*, *29*(2), 847–869. https://doi.org/10.1007/s00530-022-01025-2

Rei, L., & Mladenić, D. (2023). Detecting fine-grained emotions in literature. *Applied Sciences*, *13*(13). https://doi.org/10.3390/app13137502

Rei, L., Pita Costa, J., & Zdolšek Draksler, T. (2024). Automatic classification and visualization of text data on rare diseases. *Journal of Personalized Medicine*, *14*(5). https://doi.org/10.3390/jpm14050545

Sebastián Lozano, J., Alba Pagán, E., Martínez Roig, E., Gaitán Salvatella, M., León Muñoz, A., Sevilla Peris, J., Vernus, P., Puren, M., Rei, L., & Mladenič, D. (2023). Open access to data about silk heritage: A case study in digital information sustainability. *Sustainability*, *15*(19). https://doi.org/10.3390/su151914340

Conference Paper

Rei, L., & Mladenić, D. (2020). Cross-lingual text classification for inferring variables describing silk fabrics. *Weaving Europe Silk Heritage and digital technologies*, 177–190.

Biography

Luis Rei obtained his Master's degree in Informatics and Computing Engineering in 2011 from the Faculty of Engineering, University of Porto. His master thesis, entitled "Optimizing Simulated Humanoid Robot Skills", focused on using optimization algorithms to improve the skills of a football-playing agent. Part of this work won the best paper award at the 11th International Conference on Mobile Robots and Competitions. The thesis was a product of his experience as a member of the FCPortugal 3D RoboCup team, during which the team placed 3rd in Robocup German Open and participated in the 2010 RoboCup in Singapore.

In 2012, he joined the Sapo Labs at the University of Porto where he worked on Natural Language Processing. In 2013, he participated in the CLEF RepLab, a competitive evaluation exercise for Online Reputation Management systems, as part of Team Popstar which achieved the best results in the Filtering subtask. Later that year, Luis joined the Artificial Intelligence Laboratory of the Jožef Stefan Institute. He has been involved in a number of large European research projects in both the FP7 and Horizon 2020 programmes, including xLiMe (crossLingual crossMedia knowledge extraction), SYMPHONY (Orchestrating Information Technologies and Global Systems Science for Policy Design and Regulation of a Resilient and Sustainable Global Economy), euBusinessGraph (Enabling the European Business Graph for Innovative Data Products and Services), MIDAS (Meaningful Integration of Data, Analytics and Services), SILKNOW (Silk heritage in the Knowledge Society: from punched cards to big data, deep learning and visual / tangible simulations), ODEUROPA (Negotiating Olfactory and Sensory Experiences in Cultural Heritage Practice and Research), and enrichMyData (Enabling Data Enrichment Pipelines for AI-driven Business Products and Services).

Since 2019, Luis has also been collaborating with Event Registry, a world-leading news intelligence platform, working on information extraction from news articles, helping the company in its mission to use the power of artificial intelligence to turn news content into actionable insights.