ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DEPARTMENT OF CLASSICAL PHILOLOGY AND ITALIAN STUDIES

SECOND CYCLE DEGREE IN

DIGITAL HUMANITIES AND DIGITAL KNOWLEDGE

REVEALING CONTESTED MEMORY: AUTOMATIC SENSITIVE CONTENT DETECTION IN COLONIAL PHOTOGRAPHIC ARCHIVES

Dissertation in

Semantic Digital Libraries

Supervisor

Prof. Giovanni Colavizza

Co-Supervisor

Prof. Charles Jeurgens

Defended by

Orsola Maria Borrini

Graduation Session III

Academic Year 2022/2023

Contents

Acknowledgements

First of all, I would like to give heartfelt thanks to my supervisors, Prof. Giovanni Colavizza and Prof. Charles Jeurgens, for their support and much-appreciated advice throughout my dissertation. Without their invaluable guidance, unravelling the intricate maze at the crossroads between archival sciences and new digital technologies would not have been possible. The University of Amsterdam CREATE Lab also receives my deepest gratitude for hosting me during my study period abroad and giving me the opportunity to conduct my research in such a diverse and lively environment. I give my thanks to all involved.

I will never be able to top the acknowledgements of my Bachelor, so I will try to keep it short and sweet. Obviously, a very special thanks to my family. To my mum and dad for telling me, among many other things, to always spend money on food (especially when it's good), to my brother, for kindly providing the best stress outlet in the form of videogames, and to my cat for being itself an amazing stress outlet by simply *being*.

My other very small family in Bologna, the Righi team, also deserves a honorable mention for listening to my rambling and washing the dishes for me: thank you Costanza and Marco, *daje*!

I am very grateful for all of my friends during these chaotic two years and a half of continuous wandering: we might have not seen each other as often as we used to, but some things stay exactly the same. To Rachele, Elena P., Sara, Elisa, Schiasse, Shaula, Sampie: the coffees and aperitivi with you are always extra tasty, and the adventures even more exciting. I literally could not have done it in the Netherlands without some specific people: to Elena T. and Miriam, thank you so much for hosting and feeding me those first two weeks and for supporting my Flixbus commuting. Hopefully I will never do that again, but consider yourselves hosted back for your next adventures. Big thanks to possibly the best random discovery of 2023: Sophie, I am so glad you are not as scary as your pictures and you are actually *so fun*! I will always cherish and love our bilingual sing-alongs, hugs, stories and bike rides (and, maybe, also Angelique).

The DHDK people deserve a separate paragraph for the pure chaotic energy they bring to the table every day, sprinkled with extreme love and unwavering support, no matter the amount of debugging hours that could mean. To Fede, for being my first and last teammate, always able to provide exactly what I need exactly when I need it. To Manuele and Martina, for *sometimes thinking only about you*, the hysterical laughs at the Collegio while singing altered Disney songs and the Italian literature fangirling. To Nora, part of my little Dutch adventure, for sharing those joys and pains and for never missing the beat when it comes to singing Spirit, Notre-Dame De Paris or Hamilton. Also part of that adventure and fellow musical enthusiast, thank you Madda for the rants and for making me speculoos toasted bread in the morning, that really marked a before and after in my life. A big thanks is also due to Tom, my agent in the Netherlands, for the beers and that scooter ride, and Pablo, for always hosting the best parties and coping with the demands for a *Mambo n.5* dance. Finally, the crazy bunch: Fra, I am happy to share with you the love for worldbuilding, fiction, and also quite a bit of rage, thank you for your honesty; Manu, you have been great at enduring me all this time and you do deserve a raise, especially after saving my thesis... really, *dhanyavaad*; Ahsan, thank you for taking care of my house keys like a good landlord and for being quite literally a mischievous living meme; Chloe, *adelfi mou*, I am forever grateful for the combination of hilarious and peaceful energy you bring everywhere you go and your willingness to share it with us.

Last but not least: to Mario, because he went from explaining to me recursive functions to discussing neural networks' layers, as if I ever went beyond understanding cake fractions. That really takes lots of patience and love, thank you.

To all my friends mentioned and those that have slipped my mind, I thank you and cannot wait to keep on *ruzzando* with you.

Abstract

Although many European archival institutions hold plenty of visual materials on colonial domination, these assets are frequently difficult to access and use. Colonial records are often rendered inaccessible due to not only privacy, copyright, commercial and technical issues, but also to ethical concerns: when handling such unsettling and sensitive content, a discussion on the ethics of care and looking should be addressed, especially in relation to the digitisation of colonial archives with a focus on confronting power dynamics and amplifying underrepresented voices. The large scale of the digital archival collections originated from the digitisation efforts conducted by GLAM institutions since the 1990s marks the imperative of machine reasoning for record selection, appraisal, and management of the records. In this context, the use of Machine Learning (ML) techniques can also assist in the detection of potentially sensitive contents.

This task is investigated by constructing a full deep learning pipeline in a supervised manner: after gathering the dataset from different archival sources, the annotation is carried out by tailoring a context-specific definition of "sensitive content" and a correspondent taxonomy describing different facets of this content. The technical intricacies of utilising multi-class Image Classification algorithms for the chosen task suggest the need for a finer balance between technological efficiency and the layered analysis of colonial oppression which results in a simplification of the definition to a binary opposition between the concepts of "sensitive" and "not-sensitive" content. Several neural network architectures are then experimented and the performance of the most efficient model is validated on unseen data. The incorrect predictions are analysed by contextualising them with the developed theoretical framework. Ultimately, the results show optimistic improvements in the detection of sensitive content, albeit at the cost of heavy simplifications in the otherwise nuanced and layered subject matter.

Keywords: Archival Sciences, Sensitive content detection, Colonial archives, Machine Learning, Computer Vision, Image Classification.

List of Figures

List of Tables

Chapter 1

Introduction

The great efforts of digitisation brought forward by GLAM institutions all over the world from the 1990s turned archives in large-scale repositories of digital data. Although this paradigm shift did not initially translate into new digital archival practices, the massive amounts of digital data that were being accumulated soon led to the need for machine reasoning to select, evaluate and manage records. With the rise of Artificial Intelligence (AI) and Machine Learning (ML), its more practical counterpart, it comes as no surprise that these techniques and methods are being used in the organisation of archival workflows throughout the whole recordkeeping process [Colavizza et al., 2021].

The use of Machine Learning systems in historical colonial archives is just now emerging as a new lively field, with few but interesting and important researches such as the EyCon project (Early Conflict Photography 1890-1918 and Visual AI), which investigates the public uses of history in colonial and imperial warfare, attempting at harnessing the power of Artificial Intelligence technologies to increase the discoverability and usability of typically neglected colonial warfare materials [Aske and Giardinetti, 2023], and "Unlocking the Colonial Archive", a highly interdisciplinary project attempting at lowering the practical barriers of archival access in early modern Indigenous and Spanish collections by employing automated document transcription, text mining and automated iconography identification methods [Candela et al., 2023]. However, the postmodernist outlook on archives as active sites of contested power with harmful repercussions on underrepresented and racialised communities inevitably raises the problem of confronting such role and the possible pitfalls of the careless usage of these institutions. The simple digitisation of the records, if done outside of a decolonial effort, simply migrates the problems to a new platform which still uses traditional archival practices, considered to be heavily reflective of Western perspectives [Manžuch, 2017]. Moreover,the digitisation of colonial archives raises concerns on the ethics on large-scale viewing of potentially problematic and sensitive records and, as such, should be carefully considered [Odumosu, 2020]. Archival institutions themselves, in the last years, have attempted at enlarging their audience by breaking down barriers to access, and have done so in a push towards more diversity and inclusiveness of marginalised people in all steps of archival research and care. However, as warned by [Jeurgens and Karabinos, 2020], achieving larger inclusion "does not necessarily dismantle dominant power structures", which are likely to remain untouched. The solution proposed by these scholars is the development of modern digital archival infrastructures fostering new interactions between the historical records and modern-day users with a view to offer the multivocality, multiple agency and multiple provenance necessary for the resurfacing of long silenced voices.

Although the adoption of ML techniques can significantly ease long and tedious tasks, their high reliance on data could be the source to problematic outcomes: the algorithm, as a matter of fact, might learn the historical viewpoint expressed by traditional archival practices without being able to historicise it but rather assimilating it. Therefore, a discussion on the ethics of AI and on the ethics of care for sensitive archival records needs to be carried out before making use of these technologies on large digitised collections belonging to colonial archival institutions.

Adhering to the postmodern current of archival sciences, this thesis assesses the **feasibility of sensitive content detection in colonial photographic archives through an image classification algorithm**. "Image Classification" is a supervised learning task occupied with the automatic recognition of a set of target classes based on labeled examples used as training data. Being a supervised learning problem, its performance depends considerably on the problem definition and on the quality and dimensions of the annotated dataset. Therefore, attempting to detect sensitive images in a large digitised colonial archive first requires the development of a definition of "sensitive content" as clear as possible: given the lack of a shared definition of this concept, a new threefold description for the context of colonial photographic archives was produced along with a correspondent taxonomy defining the different facets of

colonial oppression. However, this meticulous organisation clashed with the need for simple well-defined discrete classes of a ML classification algorithm. Hence, it was simplified to a binary definition to accommodate for these technical shortcomings. The collected dataset consisted of two collections coming from different institutions home to European colonial archives: the British Imperial War Museum (IWM) and the Dutch Royal Netherlands Institute of Southeast Asian and Caribbean Studies (Koninklijk Instituut voor Taal-, Land- en Volkenkunde, KITLV). After the annotation was carried out using the Label Studio annotation tool, the labeled data was used to test several configurations of the ResNet architecture, which was chosen for its popularity and improved results tested at the ILSVRC & COCO competitions in 2015. The results prove that using image classification algorithms for the detection of sensitive content in colonial photographic archives can certainly produce encouraging results, albeit at the cost of a simplification of the definition of "sensitive content", which could entail a superficial understanding of the subject matter. Nevertheless, it is worth noting how this shallowness problem could also be overcome with the analysis of the available metadata, which has been briefly tested and proven to aid the annotation of the image content, and the use of multimodal ML models such as CLIP and GLIP, that would permit the processing of multiple modalities of data, in this case both text and image.

The thesis is structured as follows: chapter 2 describes major contributions to the field of (computational) archival studies and postcolonial studies, as well as relevant projects at the crossroad of these two fields that were necessary precursors to the development of this research. In chapter 3, the task definition problem is addressed by developing a definition for "sensitive content" and a corresponding taxonomy, used as an aid during the annotation process. Chapter 4 describes the data and methods employed and chapter 5 inspects the training experiments carried out on different neural network architectures and the error analysis of the best performing model. This model is validated and deployed to ensure that it achieves its intended purpose. Finally, Chapter 6 discusses the results and suggests possible avenues for further research, taking into account the limitations and drawbacks of this study.

Chapter 2

State of the art

The adoption of digital methods, including Artificial Intelligence (AI), in the Humanities is not a recent development, but rather a fundamental aspect of the Digital Humanities (DH) discipline.

Originally known as "humanities computing" in the 1940s and 50s, the Digital Humanities field employs digital tools to enable researchers to work with larger datasets, corpora and image archives that were previously considered unapproachable. Albeit mainly consolidated around more textual-based fields such as literary studies and linguistics, from the very beginning DH has also been concerned with audio and visual data, the latter being mainly used within art history [Greenhalgh, 2004]. Still, the visual medium has always been neglected in favour of its textual counterpart, with very little space dedicated to this field in most academic contributions: [Wevers and Smits, 2019] give a short overview of this lopsided situation and call for a turn towards the visual. The same is asserted by Taylor Arnold and Lauren Tilton, who argue that Digital Humanities research seem to be undergoing an Audio/Visual transformation due to the watershed moment brought by the combination of digitisation efforts and advances in computing in their contribution to *The Bloomsbury Handbook to the Digital Humanities*, an anthology comprising 43 essays on the key debates, methods, and possibilities of the field [Arnold and Tilton, 2022].

Cultural institutions such as GLAM (Galleries, Libraries, Archives, Museums), have been the primary contributors to digitisation efforts and have led initiatives for open access to such projects. The academic world has also funded and promoted efforts for the collection, preservation and digitisation of audiovisual materials. These materials have long been secondary objects in GLAM institutions and represent a much smaller amount of data compared to textual records. The process of digitisation is often accompanied by the annotation of records with additional contextual information and details about the digitised records. This enables disciplinary-specific analysis and precise information retrieval. Annotation is also essential in one of the most recent developments of Computer Science: Machine Learning.

Despite being used interchangeably, the terms "Artificial Intelligence" and "Machine Learning" (ML) refer to two different notions: while AI can be described as the overarching system referring to the attempt at mimicking human intelligence and human cognitive functions, ML is usually understood as a pathway to AI, a set of methods and procedures that optimise algorithms for the automation of tasks using data. The most cited formal definition of "machine learning problem" is the one provided by Tom M. Mitchell in 1997:

> A computer program is said to learn from **experience E** with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E [Mitchell, 1997].

The keyword, here, is "experience": in fact, ML involves the use of data to create and improve algorithms to perform a set task. Depending on the focus of study, AI takes on different names: particularly relevant for this project is **Computer Vision** (CV), which requires the extraction of meaningful visual information from data (digital images, videos, or other visual inputs) and the processing of that information in order to perform some task, with Image Classification and Object Detection being the most well-known. With the introduction of **Convolutional Neural Networks** (CNN) in 2012 [Krizhevsky et al., 2012], an architecture for deep learning algorithms in which the layers progressively learn more complex features of the image, the performance of CV algorithms improved drastically and the use of CNNs opened digital archives to a myriad of research possibilities, allowing the analysis of the content of historical images to discover patterns in large databases (approach mainly employed in digital art history [di Lenardo et al., 2016]), cluster visually similar records, distinguish between different types of "visual records" and perform object identification [Wevers and Smits, 2019]. A natural consequence of the adoption of deep neural networks was the belief that performance would improve more and more by adding

more and more layers. However, it was discovered that, after surpassing a certain number of layers, the performance would actually worsen: this is known as the "degradation problem". The issue was addressed by [He et al., 2016] in the paper *Deep Residual Learning for Image Recognition*, in which they introduced a new layer: the residual block, which adds a shortcut from the input features to the output of the mapping.

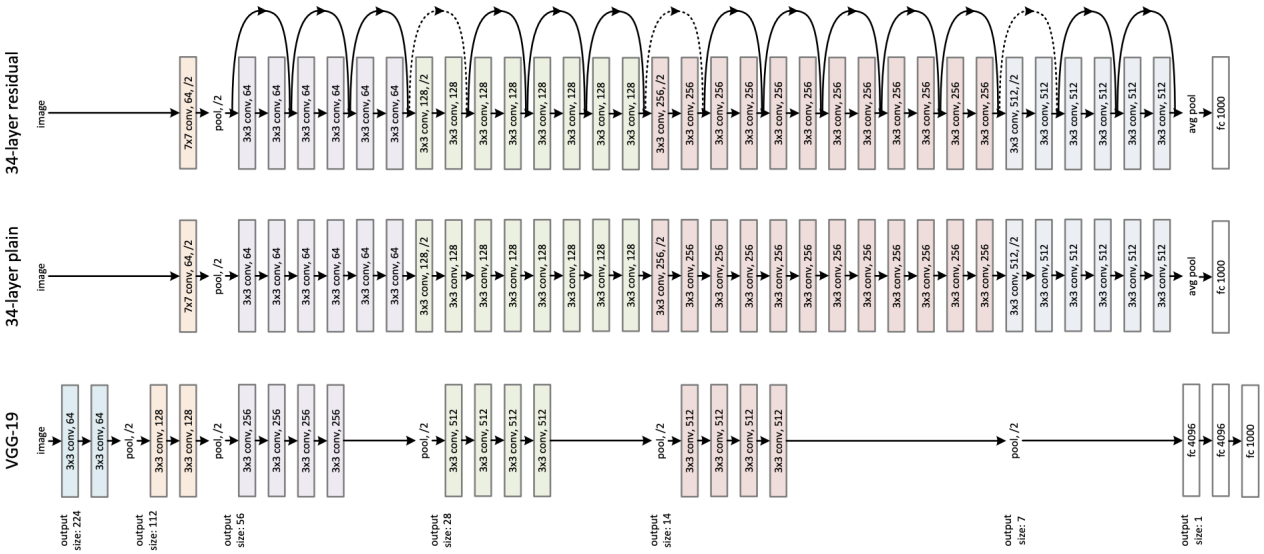

Figure 2.1: Comparison between the VGG-19 model (bottom), plain network with 34 parameter layers (middle), residual network with 34 parameter layers (top). From [He et al., 2016].

This allows for the training of networks with a large number of layers, one of the main advantages of this network for complex feature extraction, and enhances the model's robustness. The ResNet architecture, shown in Figure 2.1, was successfully adopted in a variety of different fields, from medical image analysis to image colourisation and cultural heritage applications.

There exists different types of ML approaches: while with supervised learning the algorithm is exposed to example inputs and their desired output and tries to learn the mapping function, with unsupervised learning no labeled data is provided to the algorithm, which usually then just analyses the input and tries to find patterns, clustering the data. Finally, in reinforcement learning the model interacts with a dynamic environment in a continuous attempt to maximise the provided feedback with respect to the specific task [Mohammed et al., 2016]. Supervised learning is the approach which allows for more participation on the researcher side while still being quite affordable by smaller research groups, especially since the employment of large pre-trained models, which can be adapted to similar knowledge fields through transfer learning. Moreover, supervised DNN have largely been used for several applications in the field of Computer Vision with Cultural Heritage data. Given the current status of CH large datasets, which are often unlabelled, researchers have been following the transfer learning approach for different tasks, especially in the field of art history: style, genre, and artist prediction, stroke analysis, but also archaeological remote sensing. However, DNNs are still little used by CH researchers [Fiorucci et al., 2020].

Nonetheless, there are some major limitations in using this sort of ML approach: **data annotation** can be time-consuming and expensive especially for small research groups, inconsistent due to the introduction of personal bias by human annotators and not easily scalable to larger data scopes as well as sometimes requiring domain expertise. Moreover, another important and more transversal limitation, not limited only to supervised learning, is the well-known **"black box problem"**, commonly understood as the situation in which a model is sufficiently complex that it is not easily interpretable by a human. This lack of transparency has very important **ethical consequences**, especially as AI is being used in a variety of fields which directly affect human lives: not understanding how a model makes decisions brings to several core issues including bias, fairness, and transparency. As mentioned, the data annotation phase, necessary for supervised learning, is a very delicate phase of the process, as the use of biased data has been proven to result in algorithms actively discriminating against specific groups. This has caused many concerns in the healthcare field, where explainability has been considered one of the main requirements for the application of AI [Vellido, 2019], but also has consequences in several other fields which deal with sensitive situations that might give rise to biases and ethical dilemmas such as self-driving cars and

jurisdictional tools. In fact, a lot of research conducted in the last decade brings forward the ethical and social ramifications of AI exposing the issues of bias and injustice towards minority groups in these systems [Bolukbasi et al., 2016, Caliskan et al., 2017, Lum and Isaac, 2016, Buolamwini and Gebru, 2018, Noble, 2018].

As this thesis project positions itself at the crossroads between all these topics, especially AI and (colonial) digitised archives, it was deemed fundamental to analyse and discuss the state-of-the-art situation in each of these sectors.

2.1 Access to colonial archives

The International Council on Archives (ICA) *Principles of Access to Archives* defines access as "the availability of records for consultation as a result both of legal authorisation and the existence of finding aids", specifying how one of the roles of archivists is ensuring that the records are available for use to everyone, within the "pertinent laws and the rights of individuals, creators, owners and users" [ICA, 2012]. It is clear, then, that a first possible obstacle to access is represented by privacy, copyright, commercial and technical issues.

While this is true for any sort of archive, regardless of the type of records preserved, there can be other more ethical concerns preventing or impeding the access to "archives of suffering", such as colonial archives. In general, archives have always been used by the institutions of power that created them as tools to found and corroborate a specifically curated perspective, which usually coincides with that of the Western world. This situation is further exacerbated in colonial archives, which are not only testimonies of historical periods during which racial discrimination and oppression was normalised and widely accepted in Western countries, but also working examples of such ideologies and beliefs in their own structure and functioning. Scholar Elizabeth Yakel considers archives as "representational systems" which are "both manifestations of a culture as well as the infrastructure to support that culture" [Yakel, 2003]. The common interpretation of archival institutions is the belief that they are neutral, static repositories of historical documents representing a particular society or culture. However, postmodernism has encouraged a shift in the perception of archives: from "archive-as-source" to "archive-as-subject", actively influencing the creation of historical narratives and worth investigating [Stoler, 2002]. Archives are now considered more and more as sites of power reflecting the views of their creators and distinguishing between reliable and unreliable sources, *de facto* censoring and silencing the divergent narratives.

While the most common approach to archival studies, in the last years, has been a reading "against their grain", with the proliferation of unheard stories "from the bottom up", [Stoler, 2002] argues for a parallel movement of readings, both "against" and "along the grain": to be able to adequately research the silences and presences in colonial archives, one needs first to understand the inner workings of the colonial archives, the power seeping within these mechanisms. One of the pivotal works in the discussion on power aspects of archives and their practices as power institutions is *Silencing the Past*, by Michel-Rolph Trouillot. In this book, Trouillot analyses the Haitian revolution and its reception in historical records both at the time and in modern day historical narratives. Specifically, he emphasises the materiality of history and the conflictual nature of its production. The scholar highlights four different moments in which "silences are made", underlining the deliberate nature of such occurrence: fact creation (making of *sources*), fact assembly (making of *archives*), fact retrieval (making of *narratives*), and retrospective significance (the making of *history*). The combination of these four different possible instances of "silence making" is what makes each historical narrative unique and therefore, the approach to studying it also needs to be unique [Trouillot, 2015]. Trouillot also emphasises how power is a fundamental character of historical production, a constitutional part of this process which cannot be forgotten nor put aside.

The power imbalance between "vocal" and "silenced" actors of history is further complicated by the digitisation of archives. Generally seen as a straightforward linear process, digitisation is valued as an infinite expansion of access, apart from possible technical complications, but it also conceals some problematic aspects. Firstly, the language barrier remains even in the cyberspace [Mulya and Bramantya, 2023], so much that the United Nations Educational, Scientific and Cultural Organisation (UNESCO) has long called upon institutions to promote multilingualism in cyberspace and information retrieval technologies [UNESCO, 2003]. In discussing the case of Denmark's commemorations for the centenary of the sale of the former colonial territories of the Danish West Indies to the United States, during which the Danish National Archives released relevant digitised records, [Agostinho, 2019] also argues how the digitisation of archives could support the development of decolonial practices in the archival field. However, how this process is carried out should be carefully considered as it easily lends itself to concerns about the ethics of viewing and making available sensitive and potentially problematic records on a large scale [Odumosu, 2020]: [Jeurgens and Karabinos, 2020] argue that "digitising colonial heritage is a conscious

act of (re)activation", unless the stakeholders strive to give stage to multiple perspectives and voices on the digital platforms. As already highlighted, the archival field and traditional archival practices are dominated by Western practices and this "hegemony" is usually merely transferred to the digital paradigm without really being critically addressed [Manžuch, 2017]. Moreover, even with the paradigm shift, Indigenous ways of preserving historical understanding and heritage are still discredited in favour of the host institution's needs and desires regarding the archival records under discussion [Reed, 2021, Azoulay, 2019].

2.2 AI in digitised archives

As already mentioned, Machine Learning has been used in digitised archives and GLAM institutions in general as a means to handle large corpora and to perform more in-depth research on them. Among the various surveys on the use of AI in Cultural Heritage, it is worth citing [Romein et al., 2020] for the field of digital history, and [Fiorucci et al., 2020] for a more general overview of ML used with and on Cultural Heritage.

A recent and comprehensive overview of the relationship between AI and archives is provided by [Colavizza et al., 2021], who use a recordkeeping approach to explore the role of AI in all phases of archival practice: from the creation of records, to their management and use. In particular, they employ the *theory of the Records Continuum*, a model of recordkeeping practice that conceptualises the interactions of records across interrelated dimensions [Upward et al., 2019]. With the explosion of digitisation, archives have become more and more repositories of data [Moss et al., 2018] and, as a natural consequence, newer machine and computer-based approaches and tools are increasingly being used to manage these repositories. This situation has opened up new possibilities for the role of archivists, who need to be more familiar and confident with these technologies in order to be able to deal with the new avenues of research [Rolan et al., 2019]. This has paved the way for the newly emerging field of Computational Archival Science [Marciano et al., 2018, Brown, 2018], a transdisciplinary field based on archival, information, and computational sciences, which aims to address large-scale archival research with computational methods and resources.

Typically, ML can be used to automate long, tedious workflow tasks, but its performance in this type of task has been perceived as lacking due to the significant groundwork required: not only in the area of data preparation, but also in terms of hardware requirements (local machines usually lack the necessary computational powers and the use of cloud services necessarily pertains possible privacy and legal issues). In addition, there is also a need for adequate preparation of the professionals at work: what is really needed are individuals who are skilled in both archival and computational sciences in order to make adequate use of both fields [Rolan et al., 2019]. ML is also used to improve metadata, and automated methods have been successfully used in several projects, among which there's also auto-classification for official documents [Noord et al., 2021], and to enhance the access to archives: thanks to new technologies, there can be a shift to move from the traditional access methods, based on provenance and original order, to more content-oriented access. In this sense ML is used to perform automatic content extraction and indexing and can provide new ways of accessing and researching the records [Ali et al., 2023], for example by using historical indexes' entities [Colavizza et al., 2019, Ranade, 2016] which can also pave the way to the use of Linked Data in archives [De Wilde and Hengchen, 2017].

Another experimental research strategy is that of "distant reading" [Moretti, 2013]. This approach applies computational methods to literary data in an attempt to discover "the great unread" and has also been adapted to the visual data with the name of "distant viewing" and a scope of application which includes both digital and digitised data [Arnold and Tilton, 2023]. Again, ML has also been used to search and retrieve information in large-scale archives: in the Audio and Visual data, one example is the Re:TV project, that aims to develop novel approaches for the reuse of audiovisual collections [Bocyte. and Oomen., 2020]. While ML has been has been shown to enhance inclusivity and openness in archiving by systematically maximizing the diversity of viewpoints [Gupta and Kapoor, 2020], it is still sensible to consider the unintended consequences of using AI in archives, particularly the possible biases and ethical implications.

Regarding CV applied to digitised archives, [Wevers and Smits, 2019] analysed digitised Dutch newspapers to gain a deeper understanding of visual trends. Three methods based on CNN are used to achieve this: firstly, photographs are separated from illustrations; secondly, visually similar advertisements are clustered together; finally, a network is retrained to correctly distinguish and identify different types of objects in the newspapers. The paper concludes that studying simple visual data in Dutch newspapers using only CNN highlights the need for more layered approaches that consider both the visual and textual aspects. This is in line with [Mitchell, 2005]'s concept that there is no such thing as pure "visual media", and that media should be recognized for their multiple facets and continuous two-way influence and relation. Overall, the experiments' results are clear: CNNs can semi-automatically classify large numbers of

visual sources. Additionally, they enable researchers to track visual trends, styles, and similarities over time.

On the difficulties of applying computational methods for automatic image analysis, [van Noord, 2022] has highlighted the importance of image interpretation in his inspection of the use of computational methods for iconic image analysis. In particular, he states that it is impossible to use visual-only corpora, the typical corpora in Computer Vision, to study iconic images, as for them the context is just as important as the visual content. In fact, one of the main challenges of CV is the so-called "semantic gap" [Smeulders et al., 2000] between the information that can be extracted from the visual data of the image and its significance to the user. While van Noord studies iconic images, the same problem can be perceived for what concerns colonial images, as their meaning differs depending on the context of fruition. One possible solution to this would be to develop algorithms that work in parallel with human interpretation. However, since the category of the "user" is not unique and indissoluble but rather complex and heterogeneous, it is necessary to attempt to accommodate multiple perspectives. Van Noord therefore introduces a new type of gap: the "cultural gap", i.e. the lack of correspondence between the information that can be extracted from visual data and the interpretations that the same data has for cultural groups over time. Like [Wevers and Smits, 2019], even van Noord calls for a complementary approach that uses other modalities to get a full interpretation of meaning, rather than just the visual information. He also addresses problems such as the lack of connections between the collections of different institutions, notwithstanding efforts such as Linked Open Data [Marden et al., 2013] and the International Interoperability Framework (IIIF) [Snydman et al., 2015], and argues that there is still no accessible large-scale dataset of visual culture from different time periods and domains, obviously also due to the need for datasets to meet ethical standards [Prabhu and Birhane, 2020]. This clearly means that each project must, to some extent, build its own dataset for its own specific needs, making it more expensive to fund and conduct this kind of research.

2.3 AI in colonial archives

Although not numerous, there are some important projects applying AI technologies and techniques to digitised archival collections. These projects represent a significant step forward in leveraging ML to enhance the management and accessibility of digitised archival collections. By harnessing the power of Artificial Intelligence, these initiatives aim to revolutionise the way we preserve and explore historical artifacts and documents.

[Luthra et al., 2022] use automated entity recognition to give voice back to the silenced marginalised people of the VOC Archive, the archives of the Dutch East India Company. This research stems from the role archivists play in facilitating access to the content of records: by participating in a supposedly neutral and impartial process, they have reproduced the dominant view and practice that clearly facilitates certain identities and people over others. However, even when they have attempted to make access to records multifaceted through indexing, the choice of who to index has ended up reinforcing a discriminatory approach that again privileges certain groups. In order to carry out a truly open indexing of colonial archives, one that does not replicate the oppression and discrimination of the past, the work presented proposes the use of automated entity recognition in an attempt to surface the mentions of people hidden in the documents. The task of named entity recognition in historical records is usually performed by neural network approaches, in particular by fine-tuning large pre-trained embeddings, but the importance of large datasets, low levels of noise and pre-cleaning phases are still of central importance. The results of this research are encouraging and entice further research in this area. Nevertheless, the authors acknowledge the potential problems of the proposed approach and avoid interpretations of group identity, refraining from attempting to identify people's origins as this automatically raises questions about ethnicity and its representation. They also call for ongoing dialogue between all parties and stakeholders to find the right balance for respectful access to colonial archives.

The interdisciplinary project *Unlocking the Colonial Archive* attempts to solve the problem of access to colonial archives by using ML technologies to transform the intellectual inaccessibility of Spanish American documents into open, accessible data. Specifically, it investigates three different areas: Handwritten Text Recognition (HTR) to automatically transcribe Spanish American documents; NLP techniques and Linked Open Data models to identify and link information within these corpora; Computer Vision approaches to facilitate the automated search and analysis of pictorial elements. The project is funded by the AHRC, UK Research and Innovation (UKRI), and the US National Endowment for the Humanities (NEH) and it is carried out with the collaboration of the University of Texas, Lancaster University, and the Liverpool John Moores University. Part of it is described by [Candela et al., 2023], who attempts at transforming digital collections into Linked Open Data following best practices. Specifically, this work seeks to transform Indigenous and Spanish colonial archives from the *Relaciones Geograficas de Nueva*

España collection into readable and accessible data using ML technologies, encouraging CH organisations to adopt LOD in their collections through the application of Semantic Web technologies, hence providing wider access. The work carried out in this research is particularly relevant for the Computer Vision experiments used to detect key iconography that describes cultural elements and visual landmarks, such as those found in map paintings. The authors and contributors recognise that, despite modern innovations, the majority of technologies used to describe knowledge are still rooted in Western worldviews and dominated by a white, patriarchal perspective [Risam, 2018, D'ignazio and Klein, 2023], and suggest including counter and alternative narratives, such as those coming from the Global South, in AI and other computational tools to prevent the invisibility of a wide range of knowledge that has been overshadowed by the preference for Western rationality as the only framework for existence, analysis, and thought [Mignolo and Walsh, 2018]. This can be achieved by engaging with data objects in new ways and through new forms of knowledge.

[Aske and Giardinetti, 2023] discuss the EyCon project, funded by AHRC/LABEX and focusing on Early Conflict Photography from 1890-1918 and Visual AI. The project, presented as a continuation of the research conducted by the ADDI project in object-based computer vision techniques to enhance image metadata, explores the potential for ML to improve metadata creation and enrichment. Currently, digitised visual data often lacks important metadata and the existing one is often problematic due to outdated terminology or inconsistent information. Digitised photographic collections often have issues such as missing information (i.e., on the origins, donors, acquisition details, commentaries, and institutional histories of the records), and their dispersal across various GLAM institutions, which can lead to issues with accessibility, incomplete cataloguing, and the invisibility of some archival records. Digitisation processes have made missing or incorrect information more visible. However, this process does not always generate precise metadata, which affects the accessibility and discoverability of digitised visual records, especially in large-scale datasets and collections, especially if they are mainly searchable through the use of keywords. This problem is particularly acute for potentially problematic or sensitive materials, which cannot benefit from bespoke metadata given the scale of digitisation processes.

Furthermore, metadata can be problematic as it can affect the "interpretation, exploration and use" of visual records, creating space for further distortion, misinterpretation, misappropriation and biased narratives of historical events [Verstockt et al., 2018]. Issues of missing or incorrect metadata and contextual information inevitably affect the current ML algorithms used in these cultural institutions. While Ey-Con focuses mainly on the automatic creation of metadata to counteract this problem, it also addresses another fundamental issue: the presence of sensitive content, which is inevitable given the source of the collections. The exploration and research of CVs in CH institutions has mainly focused on artworks, but the application of ML models has also played a valuable role in increasing the accessibility and usability of image archives; in archival institutions, these tools are largely pre-trained on colour images of Euro-American visual culture materials such as COCO and ImageNet, and as such have little to no overlap with historical records. In fact, many existing training datasets have been shown to introduce forward bias [Smits and Wevers, 2022]. As Lise Jaillant has argued, the application of AI to archives risks permanently affecting our collective memory and distorting the historical record. And when these automated systems are used with potentially sensitive or controversial materials, the ethical implications are compounded [Odumosu, 2020].

The EyCon project aims to develop interoperable models to create a pipeline for historical semantics research. One of the first obstacles identified was how to deal with problematic metadata: this issue is usually addressed by cultural institutions through the use of policies and statements published by the institutions, but there is no common set of guidelines that specify how to deal with this problem, even more so when metadata is itself a historical resource. This is the belief of the EyCon project, which argues that outdated or inappropriate metadata should not be removed, but rather preserved and further explained. However, it is emphasised that this should always be done in collaboration and in parallel with more traditional human-generated metadata, particularly given the sensitive nature of the subject matter. To avoid bias, images are categorised into larger categories and controlled vocabularies are used to match existing image archive ontologies. One of the most interesting aspects of the EyCon project is the identification of sensitive content. While acknowledging the subjectivity of sensitivity, they have used object detection to help identify sensitive content experimenting with Google Vision's "Detect Explicit content - Safe search" to classify the photographs of the corpus with more graphic content (such as corpses); while this has yielded poor results, further experiments are being carried out using pre-trained models that are re-trained on high-quality annotations completed by experts from a broad consortium of institutions in the GLAM sector.

Chapter 3

The definition of sensitive content

3.1 Research problem

One of the first steps in Machine Learning (ML) is the **problem definition**: after determining which task the model is going to be trained for, its description is a fundamental prerequisite to performing coherent data annotation and reach a better performance. In this research, the selected task is the detection of sensitive content in the specific context of colonial photographic archives and, therefore, a clear definition of "sensitive content" should be established.

The notion of "sensitive content" does not have a fixed and shared meaning. Its definition changes depending on the purview of inquiry: what is considered "sensitive" in social media is not necessarily regarded as such in financial, legal, or healthcare data. Cultural heritage institutions such as GLAM (Galleries, Libraries, Archives, Museums), generally provide cautionary statements to alert the users of the presence of potentially sensitive content in their collections to allow them to decide whether to engage or disengage with the records. These notices come in all shapes and forms but, as analysed by Lindsay Loebig, they usually do share some common features: they provide context to the presence of possibly hurtful content, encourage users to submit feedback, acknowledge the presence of bias in the institutional praxis, and cite and share resources [Loebig, 2023]. Usually, the notice is also accompanied by a statement highlighting the outdated nature of such content and how the position expressed does not align with the current values and practices of the hosting institution.

By comparing the statements published by different institutions, we can start outlining a very generic and broad definition of sensitive content: anything that may be harmful, offensive, or misrepresenting especially when in relation to religion, race, gender, politics, sexuality, disability.

For this project, it was deemed sensible to provide a **context-specific definition** constructed on the analysis of three main aspects: colonialism, photography, and archival institutions.

1. **Colonialism**
   (a) Even the phenomenon of "colonialism" lacks a universal definition and it is generally classified depending on its specific goals and assets (some examples being settler colonialism, exploitation colonialism, and surrogate colonialism). In this work, we use Jurgen Osterhammel's three-sentence definition:

   > "Colonialism is a relationship between an Indigenous (or forcibly imported) majority and a minority of foreign invaders. The fundamental decisions affecting the lives of the colonised people are made and implemented by the colonial rulers in pursuit of interests that are often defined in a distant metropolis. Rejecting cultural compromises with the colonised population, the colonisers are convinced of their own superiority and their ordained mandate to rule." [Osterhammel, 2005]

   (b) The data used for this research is sourced from European colonial archives: consequently, it is important to pay attention to the ideology of "race", one of the major driving forces behind European colonialism. We will discuss this problematic aspect further on, in subsection The importance of "race" and colour blindness.
2. **Photography**
   (a) As a record keeping technique, photography inherently possesses problematic features: in her pivotal work *On Photography*, Susan Sontag points out how, it being an interpretation rather than an accurate depiction of reality, photography easily lends itself to becoming a

tool of power through which people take possession of a space in which they feel insecure [Sontag, 1979]. This power is also evident in the act of gazing, which is made permanent and therefore ever impactful by the production of photographic records and their circulation. When researching sensitive photographic records, Susan Crane alerts on their mindless consumption which posits the risk of reiterating the violence through which they were produced in the first place [Crane, 2008].

(b) Since its invention, it became apparent how photography would be the perfect instrument in administrative, missionary, scientific, and commercial activities in the colonial dominion [Cole, 2019]: as a matter of fact, it was profusely used to produce an ordered, well-rehearsed, curated narrative constructed specifically for the European audience [Foliard, 2021] and further fortifying opposing subjectivities [Odumosu, 2020] (the binary opposition of Self/Other pointed out by Said in *Orientalism* which took a prominent place in postcolonial studies [Said, 1995]).

3. **Archival institutions**

(a) Archives, commonly understood as institutions that accumulate historical records or materials, are now being considered sites of power and narrative creation: through the selection, preservation, and access to the records, they wield that power, creating a differential ranking of narratives and perspectives and establishing "what is remembered and what is forgotten, who in society is visible and who remains invisible, who has a voice and who does not." [Macknight, 2011]. As Michel-Rolph Trouillot states in *Silencing the Past*, these presences and absences are not neutral or natural, "but mentions or silences of various kinds and degrees. By silence, I mean an active and transitive process: one 'silences' a fact or an individual as a silencer silences a gun." [Trouillot, 2015].

(b) Archival institutions and systems are still highly dominated by Western perspectives and little space is usually given to diverse cultural representations and narratives: it is under this Western paradigm and its "objective" and "neutral" gaze that the Indigenous *Other* was created [Anderson, 2013] and the colonial power was incessantly reinforced and legitimated [Macias, 2016]. As argued by Susan Crane for atrocity photographs, their consumption inevitably means taking the perspective of the perpetrator and, therefore, participating in the atrocity [Crane, 2008]: through this lens, treating colonial archives' records as simple data would mean refusing to acknowledge and further contribute to and deepen the living, multigenerational trauma [Subotić, 2020] these materials are representation of.

3.1.1 Sensitive content *for whom?*

The scope of this research required the development of a clear definition of "sensitive content". However, it must be stressed that this is a very subjective matter, especially when dealing with such delicate topics which still, to this day, have an impact on the Indigenous communities affected. The careless reproduction and consumption of materials coming from European colonial archives within the traditional archival practice runs the high risk of perpetuating the hurtful, violent, and discriminatory views in which these records were produced. Given the historical context, the "sensitivity" of these materials targets the Indigenous communities affected by colonial dominions.

The proposed definition of "sensitive content" is not an attempt at ending the debate, but rather a call for greater inclusion of these communities in the whole workflow of archival practice and research to proceed with a process of decolonization of these record keeping systems, too often falling back into a pervasive and damaging inaction with respect to their problematic colonial records, through the facilitation of multiple perspectives and the opening of the system to communities to facilitate the desperately needed multivocality [Jeurgens and Karabinos, 2020].

Without the inclusion of the communities which have been affected by the system author of these records in the first place , we cannot lightly define "sensitive content", depriving them, once again, of their voice. In this sense, when using the term "sensitive" here, it is to be understood as a synonym of "delicate": these are archival materials which need to be handled with additional care: such care goes well beyond generic content notices and must give rise to a process of decolonization.

Moreover, this involvement should also allow for the removal and potential destruction of Indigenous cultural records. Trevor Reed argues that "for Indigenous peoples to retain and fully exercise their sovereignty, they must have the power to create culture, but also, when necessary, allow that culture to die." [Reed, 2021], especially considering that most of these records, if not all, have been created under ethically dubious circumstances. Prioritising archival procedures instead of Indigenous communities'

modes of caring means exercising the same violence with which these records were appropriated or created [Azoulay, 2019].

3.2 Premises and limitations

Considering all the above, a few premises and limitations should be made:

1. All records held in colonial archives should be considered sensitive due to the oppressive circumstances in which they were produced.
2. In this research, the detection of sensitive content is performed only on the visual content of the image, leaving out the metadata and contextual information in the interests of simplification.
    (a) This implies that the intersectional nature of oppression and other possible gender and socio-economic analyses have been overlooked in favour of the colonial aspect.

3.3 Definition

The delineation of a definition was heavily influenced by the data analysis which was conducted in parallel: from the examination of the pictures available it quickly became apparent how, as a natural consequence of premise 1), rather than giving a comprehensive understanding of the notion of "sensitive content", it would have been more appropriate to aim at describing **different degrees of recognisability** of sensitive content.

Therefore, the proposed definition is structured as an ordered list of recognisability degrees in a descending order (from most to least recognizable): sensitive, dubious, not-sensitive.

1. **Sensitive content**: content which is more explicit and easily recognisable as immediately sensitive.
    (a) Any document reiterating the colonial racist, xenophobic, and discriminatory beliefs, providing harmful and stereotypical representation of Indigenous individuals and communities and demonstrating, in general, bias and exclusion of marginalized people.
    (b) Any document depicting explicit violent graphic content including death, medical procedures, crimes against the person, war, and terrorist acts.
    (c) Any document containing clear symbols and references to the colonial context (e.g., foreign invaders' flags, portraits of the rulers, medals...).
2. **Dubious content**: unclear content which would benefit the most from the contribution of Indigenous communities and experts to the workflow.
    (a) Any document that could be triggering or cause offence to the Indigenous communities affected by colonial imperialism if not presented with the due consideration and empathy (e.g., studio portraits, Cultural Heritage with people...).
3. **Not-sensitive content**: content which does not display any clear or explicitly sensitive feature (e.g., crowd pictures, landscapes, catalog pictures of Cultural Heritage and common objects...).

Each of these degrees corresponds to a class during the annotation phase. However, due to the need to find a compromise between a full in-depth theoretical representation of such a complex problem and the need for a clean distinction between classes to achieve the best possible model performance in Machine Learning, this definition was simplified by distinguishing two separate classes: "sensitive", incorporating both "sensitive content" and "dubious content", and "not-sensitive". This allowed a first attempt at classification using **binary classification**, which is structurally simpler but more robust and more likely to give better results. At the later stage of error analysis and results interpretation, the problem was returned to its original three-way structure and tested again.

Some examples are shown in figs. 3.1a to 3.1c, 3.2a to 3.2c and 3.3a to 3.3c.

(a) Leiden University Library, KITLV 19037

(b) Leiden University Library, KITLV 6295

(c) Leiden University Library, KITLV 19030

Figure 3.1: Examples of "sensitive content"

(a) Leiden University Library, KITLV 32139

(b) Leiden University Library, KITLV 32092

(c) Leiden University Library, KITLV 92610

Figure 3.2: Examples of "dubious content"

(a) Leiden University Library, KITLV 66765

(b) Leiden University Library, KITLV 65589

(c) Leiden University Library, KITLV 92755

Figure 3.3: Examples of "not-sensitive content"

3.3.1 Issues and shortcomings

Unclear classes' boundaries

As can be seen from the provided examples, it is clear that the three different classes are not always easily distinguishable from one another.

This is particularly evident in the case of "sensitive" and "dubious" content, which can be mixed up especially when the annotation is mainly dependent on **symbol detection**. The use of this criterion to distinguish between borderline images between the "sensitive" and "dubious" classes could be regarded as problematic and weak for different reasons: the simple detection of these triggers in the images can encounter several technical barriers in an insufficient file quality or in the colour mode of the picture, which could make it impossible to distinguish these symbols from a busy background. During the annotation these details were considered sufficient to distinguish between a "dubious" and a "sensitive" label for images which shared all or most other phenotypic features, it is important to note how this could be the one of the causes of the fragility of classes' boundaries.

One example is shown in Figure 3.4, where both instances are examples of studio portraiture, a genre associated with the dubious category in the proposed classification, but have different labels: while (a) is annotated as having "dubious content", (b) is considered to have "sensitive content" due to the presence of the medals, considered symbols and referent of the colonial oppression in place at the time.

(a) Leiden University Library, KITLV 32143

(b) Leiden University Library, KITLV 7834

Figure 3.4: Examples of different annotations based on symbol detection

Lack of depth

The definition lacks the necessary depth to understand the pervasive nature of the colonial occupation: in a colonial dominion, all aspects of Indigenous lives were entirely regulated by the occupants. This less evident oppression, still with real and hurtful consequences, is not adequately understood in the proposed definition and risks falling under the hood of the "dubious content".

More extensive knowledge and expertise in both historical and cultural aspects would certainly improve the overall outcome of the research: Indigenous communities' and experts' inputs are fundamental.

3.4 Taxonomy development

The complexity of the problem definition phase, heavily influenced by the intricacies emerged during the data analysis, was tackled with a definition built on a delicate balance between the different possibly sensitive features of each image and the need of clarity and structure for the Machine Learning task. However, the data annotation phase still needed an additional aid to carry out the task as homogeneously as possible.

As a matter of fact, providing an precise and consistent annotation of the data is crucial to achieve an accurate performance of the model and a pilot of this phase is sometimes carried out in parallel with the task definition, to allow for flexibility and possible resolutions in the problem definition as well.

For this reason, a taxonomy of sensitive content was produced as the result of both continuous discussion on the available material and the current literature on racial discrimination and European colonialism such as, among others, the U.S. National Research Council's analysis of racial discrimination [Council et al., 2004], and Edward Said's pivotal work *Orientalism*.

The taxonomy, a classification scheme best known for its use in biology, is an empirically-based categorisation of individual samples which are grouped based on their different observable features (i.e., their phenotype) [Bailey, 1994] and can be implemented in a variety of different disciplines. In particular, a mix of the empirical-to-conceptual and conceptual-to-empirical design approaches were used, analysing both the data and the literature on the topic phenomenon [Kundisch et al., 2021].

3.4.1 Approach

The development of the taxonomy was carried out adapting the procedure employed by Paul Brecke for the purpose of conflict early warning [Brecke, 1997] and it can be summarised in the following steps:

1. Define the population of data to be categorised and collect a sample.
    - In this research, the data is all the pictures with sensitive content from colonial photographic archives, and the sample is the selection of records forming the dataset (see Section 4.1)
2. Define the set of variables describing the data.
    - Specifically, this step was completed after a first cycle of data analysis during which the triggering features for each record were noted down and discussed, particularly for what concerns problematic instances (specifically, the recognition of different "races" or "ethnicities" and the problem of "colour blindness").

The importance of "race" and colour blindness

During the data analysis it became clear that, among the different phenotypic features detected as plausible triggers for "sensitive content", there was a possibly problematic one: the recognition of different "ethnicities" or "races" [1].
Although most scholarly fields now accept the concept of "race" as a social construction with no real scientific support, it is undeniable how its definition has been used for purposes of domination and has facilitated a process of "othering", with definite social consequences. Moreover, Michael Omi and Howard Winant have asserted that, although "race" implicates multiple variables such as language, culture, and religion, racial distinctions are based on a set of phenotypic characteristics and this visual dimension is crucial to the distinction between racial categories [Omi and Winant, 2014].

However, the use of this feature in the annotation phase easily runs the risk of producing a racially biased ML model. This has been a key concern regarding the development of Machine Learning models especially with Computer Vision systems, which highly depend on the training datasets. There has been consistent evidence of racially biased Artificial Intelligence algorithms with important real-world consequences. Some examples include Google Vision Cloud, which labelled the same image differently based on the skin colour of the person present, and facial recognition technologies which have been found to being significantly less accurate for people with darker skin tones by a study of the U.S. National Institute of Standards and Technology [Schwartz et al., 2022].

[1] For more information on the choices taken regarding the use of language, see Appendix A.

Given all of the above, "race" was not considered as a phenotypic category in this research. Nevertheless, it is important to stress the real consequences racial oppression has had and still has on different oppressed groups (removal, genocide, slavery, invasion, colonisation and exclusion being only some of them) and the shortcomings of a colorblind view of "race": the solution is not ignoring "race", but acknowledging how it still, to this day, generates inequality. To address these racial inequalities, race-conscious policies and practices should be applied [Omi and Winant, 2014]. In the case of academic archival research, we believe that this translates in a wider inclusion of Indigenous community for the application of a decolonising practice of the traditional system.

3. Based on the phenotypic features, define a set of abstract categories and their possible values.

   (a) Clothing style
   - Only one
   - Various

   Caveat! Given the colonial context, it could be possible for the "clothing style" to end up becoming a proxy for the unused "race" category. While it is not possible to neglect the "clothing style" category as well, it is still important to highlight this issue.

   (b) Pose of the subject(s)
   - Different poses
   - Posed

   (c) Action type
   - Direct abuse
   - Indirect abuse (e.g., crying, scars...)

   (d) Background
   - Blank
   - Artificial

Quite soon during the annotation, it became clear how what constitutes a trigger for the detection of "sensitive content" in its different degrees is not just the presence of one of these categories, but rather the combination and coexistence of multiple categories at the same time. In particular, a set of combinations was considered the most relevant, as shown in table 3.1.

Clothing style		**Pose of the subject(s)**		**Action type**		**Background**	
Only one	Various	Different poses	Posed	Direct abuse	Indirect abuse	Blank	Artificial
		(•)		•			
		(•)			•		
	•		•			•	
	•		•				
•			•				
	•	•					

Table 3.1: Categories combinations

3.4.2 Taxonomy

The produced taxonomy only applies to the data of the categories "sensitive content" and "dubious content", as it aims at describing different types of more explicit sensitivity, and considers four different taxa: Violence and abuse, "Scientific" racism, Otherness and Exclusion and segregation. For each of these units, different components are indicated (and more could be added with further research): Violence and Abuse can be either explicit or implicit; "Scientific" racism can be represented as physical anthropology; Otherness can appear as exoticism in the form of fetishisation; Exclusion and segregation are present in the form of either statistical or structural discrimination. The taxonomy is presented in figure 3.5.

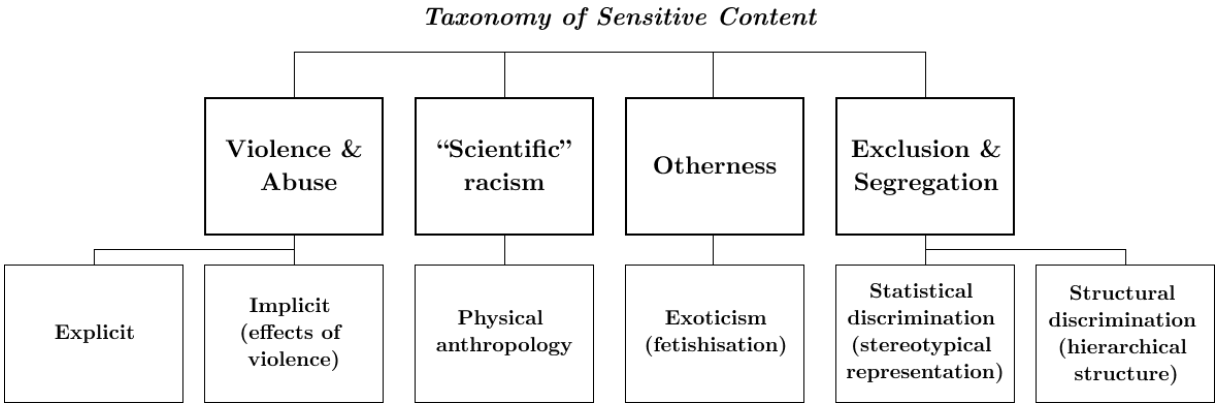

Figure 3.5: Taxonomy

The correspondence between the combinations of categories and the taxonomy classes is shown in table 3.2.

Clothing style		**Pose of the subject(s)**		**Action type**		**Background**		**Taxonomy**
Only one	Various	Different poses	Posed	Direct abuse	Indirect abuse	Blank	Artificial	
		(•)		•				Violence and abuse (explicit)
		(•)			•			Violence and abuse (implicit)
	•		•			•		"Scientific" racism
	•		•					Otherness
•			•					Exclusion and segregation (statistical)
	•	•						Exclusion and segregation (structural)

Table 3.2: Taxonomy and categories correspondence

Chapter 4

Data and methods

4.1 Data

The data used in this research was collected from two different archival institutions: the British **Imperial War Museum** (IWM), and the **Royal Netherlands Institute of Southeast Asian and Caribbean Studies** (Koninklijk Instituut voor Taal-, Land- en Volkenkunde, KITLV).

The IWM collection consists of a sample of 199 images that were previously prepared for digital work during the EyCon project (Early Conflict Photography 1890-1918 and Visual AI): these photographs are digitized to a high-quality standard and focus on several colonial wars including several theaters in World War I and the Boer Wars. Although metadata was provided, it was not used in the research as it only pertains to the visual content of the images.

On the other hand, the KITLV collection was compiled using **webscraping techniques** from two digital libraries, resulting in a total of 2336 images. The initial webscraping was performed on the image bank "Het geheugen van Nederland" (GvN), a service of the Koninklijke Bibliotheek. Although no longer updated, GvN still contains a significant amount of data, particularly from the KITLV. However, during the annotation process it became clear that many files did not meet the project's quality standards: in wide shots, where the subject(s) are captured from a distance, details were often lost when zooming in, making it difficult to identify potential triggers for annotation. This webscraped collection ("The Dutch East Indies in photographs, 1860-1940") belongs to the KITLV, which has been housed at Leiden University Libraries since 2014 as part of the KITLV library collection. Therefore, the scraping was carried out again on the Leiden University Libraries' Digital Collections website, filtering the KITLV as needed (using the keywords: "Nederlands-Indië in foto's, 1860-1940"). Unfortunately, this resulted in a lower number of pictures and the best quality pictures from the "Het geheugen van Nederland" were therefore kept as part of the dataset to increase its size, already very small for a Computer Vision task.

The data came in various formats depending on the source, with PNG (Portable Network Graphic) files obtained from the IWM and GvN collections and TIFF (Tag Image File Format) files from the Leiden University Libraries' "Nederlands-Indië in foto's, 1860-1940" (ULNI) collection.

Institution	**Number of pictures**	**File format**	**ID**
Imperial War Museum	199	PNG	IWM
Het geheugen van Nederland (KITLV)	355	PNG	GvN
Leiden University Libraries' Digital Collections (KITLV)	1978	TIFF	ULNI
Total	2177 (33,1 GB)		

Table 4.1: Dataset composition

4.1.1 Terms of use

As stated in their Copyright section on the website, "Het geheugen van Nederland" is a web service of the Koninklijke Bibliotheek, to which the Copyright and the Database Act apply. Specifically, the pictures used for this project and here reproduced belong to the KITLV, which is the referent to contact for copyright information. The rights status of all the KITLV resources presented in this research is **public domain** and the pictures are cited according to the guidelines specified by the Leiden University Libraries. No pictures from the Imperial War Museum have been included, as they have not yet been published.

4.2 Methods

4.2.1 Annotation

The raw data was annotated using **Label Studio**, an open source data labeling tool for labeling, annotating, and exploring a variety of data types. Each of the three collections was uploaded as an independent project for organisational purposes and to partly overcome the low storage space problem encountered when uploading files: in Label Studio, the default maximum total size of all files uploaded from a local directory is set to ≈260 MB, which is definitely too little if compared to the size of the entire dataset used in the project: 33,1 GB.

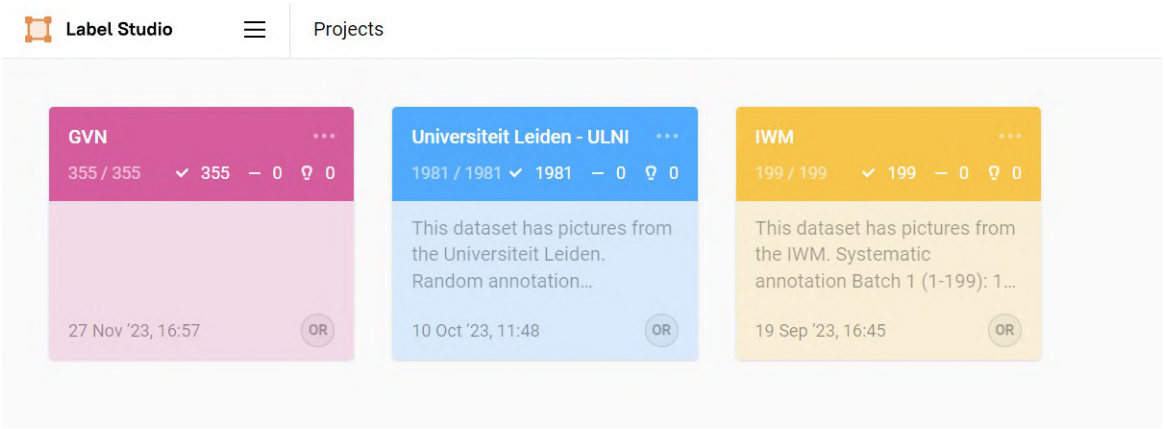

Figure 4.1: Screenshot from the Label Studio User Interface

Moreover, Label Studio does not support the TIFF format for the input data (the supported file types for images are .bmp, .gif, .jpg, .png, .svg, .webp), and the images from the ULNI collection, the largest one in the dataset, had to be converted. The batch conversion to the PNG format was done using **IrfanView**, a freeware graphics viewer for Windows. Finally, as this collection is also the largest in the dataset at 32,6 GB, it was uploaded via a **bash script** running a web server to generate URLs to the local files.

The labeling interface was created by adapting the "Image Classification" template to the needs of the project, configuring the three classes of "sensitive content", "not-sensitive content" and "dubious content", and the annotation was performed using the previously developed taxonomy by randomly sampling the tasks to support an annotation approach as homogeneous as possible.

Upon completion of the annotation process, the labeled data was exported in CSV format, with the structure shown in Table 4.2. Each collection was uploaded as an independent project, resulting in three distinct CSV files.

annotation_id	annotator	choice	created_at	id	image	lead_time	updated_at
1	1	Sensitive content	2023-09-19 T14:53:45.965682Z	1	/data/upload/1/ 6ba74cc4-Q_017042.jpg	44.225	2023-09-19 T14:53:45.965682Z
...	...	...	...	...	...	...	...
4	1	Dubious content	2023-09-19 T14:55:19.854238Z	4	/data/upload/1/ 613fc759-Q_017045.jpg	100.020	2023-10-03 T11:00:38.615262Z
5	1	Not-sensitive content	2023-09-19 T14:55:26.779273Z	5	/data/upload/1/ 9cdb6487-Q_017046.jpg	6.623	2023-09-19 T14:55:26.779273Z

Table 4.2: Structure of the annotated data (CSV file)

4.2.2 Data preparation

The annotated data then went through a number of different processing stages: these and the other procedures described in the following chapters are described in more detail in the documentation uploaded to the public GitHub repository [Borrini, 2024] in the form of four **Jupyter Notebook** files.

The first step was that of **data cleaning**: unnecessary or redundant information was deleted for clarity, while new information about the original collection was added to keep track of it in further steps. After several problems were encountered during the training, an additional step was added in this phase: the deletion of damaged images (which raised alerts as "truncated" files) and, consequently,

the modification of the CSVs to take into account the disappearance of these files. Another issue that arose during the training was the different number of channels in the images: while the IWM images are grayscale (one channel), the GvN and ULNI images are in RGB (three channels). The images were therefore all converted to the RGB "mode" using the PIL Python library.

Finally, all of the images were manually temporarily moved to a common folder and the three CSV files were merged into an index file to keep track of all the available information on the dataset, and the image path information was updated with the temporary one.

The index file is presented in Table 4.3.

annotation_id	choice	created_at	id	updated_at	provenance	set	image
1969	not-sensitive content	2023-10-18 T09:47:11.226933Z	5192	2023-10-18 T09:47:11.226933Z	ULNI	NaN	pictures/ UL_NI_812.png
...	...	...	...	...	...	...	...
2736	dubious content	2023-11-27 T16:06:41.267915Z	6975	2023-11-27 T16:06:41.267915Z	GvN	NaN	pictures/ GVN_237.jpg
115	sensitive content	2023-09-26 T15:38:39.340779Z	115	2023-09-26 T15:38:39.340779Z	IWM	NaN	pictures/ Q_052373.jpg

Table 4.3: Structure of the cleaned data (CSV file)

The last step before the experimental training was the **creation of the dataset** to enable model training, performance evaluation, and publication through testing on previously unseen samples. The available data was split into three sets: train, validation, and test. Typically, the train set has a larger number of samples than the validation and test sets. In this project, we split the data in a 70-15-15 ratio. Although randomisation is sometimes used in this procedure, a quick data analysis revealed a significant imbalance in the dataset, with a bias towards the "not-sensitive content" class, as shown in Table 4.4. It is important to address this imbalance to ensure accurate results.

Class	Samples	Percentage
not-sensitive content	1939	76.58
dubious content	330	13.03
sensitive content	263	10.39

Table 4.4: Classes proportions in the dataset

Using a random split on such an imbalanced dataset would risk producing a train set where one or both of the least abundant classes do not appear. This would cause the model to fail to learn the features of the missing class or classes, resulting in an unreliable performance of the model. To prevent this issue, the data split was done through **stratified random sampling**, ensuring the persistence of the dataset population classes' proportions, despite unbalanced, in all three sets.

The Table 4.5 presents the overall dimensions of the sets.

Set	Number of samples
Train set	1772
Validation set	380
Test set	380

Table 4.5: Sets' dimensions

Finally, the dataset is structured in folders where the images are organised based on their set and class, as seen in Figure 4.2, and the index CSV file is updated with the new file paths and set values.

4.2.3 Model training

The dataset thus created was used to experimentally train different pre-trained model configurations in order to find the best hyperparameter setup for the "Image Classification" task. The best performing model was then selected and validated. Finally, its performance was analysed to understand the possible sources of errors and failures in the predictions.

This phase was carried out mainly using **Hugging Face** libraries (such as `datasets`, `transformers`, and `evaluate`), which supports the interoperability between frameworks and allow easy download and

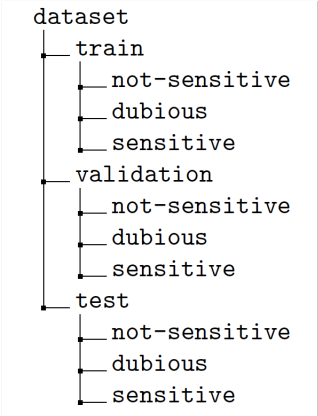

Figure 4.2: Dataset folder-based structure

training of state-of-the-art pre-trained models. Hugging Face is a machine learning and data science platform and community that aims to democratise artificial intelligence for all: they work towards this goal by providing users with tools to build, train, and deploy ML models.

Moreover, the experiments were tracked with **Weights & Biases**, an AI developer platform with tools for training, fine-tuning, and leveraging foundation models. The tests can be examined on the public dashboard, accessible from the Github public repository dedicated to this project [Borrini, 2024].

The adopted architecture for this research is a Residual Neural Network; **ResNet** is a Convolutional Neural Network first proposed in the paper *Deep Residual Learning for Image Recognition* [He et al., 2016] as an improvement over the traditional deep networks by introducing residual connections that facilitate the training of deeper neural networks with an unseen number of layers. Specifically, the version used is v1.5, **ResNet50**, developed by Nvidia by modifying the original model: the difference between v1 and v1.5 is that in the bottleneck blocks that require downsampling, v1 has stride=2 in the first 1x1 convolution; whereas v1.5 has stride=2 in the 3x3 convolution. This difference means an improvement in accuracy for ResNet50, albeit with a small performance penalty. When used within the Hugging Face framework, this model is loaded with an image classification head on top (a linear layer on top of the pooled features). The checkpoint used in this research is "microsoft/resnet-50", pre-trained on ImageNet-1k at a resolution of 224x224.

The choice for this architecture instead of the newer Visual Transformer (ViT), which was proposed in *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale* by [Dosovitskiy et al., 2020] and adopts the Transformer's attention mechanism [Vaswani et al., 2017] to more easily approach Computer Vision tasks, was due to the poor performance of ViT with small datasets compared to traditional CNNs. [Zhu et al., 2023] explore this difference in performance by comparing the ViT and ResNet architectures on three different datasets and speculate that the reason of the failing ViT model is the lack of inductive bias of locality, regardless of the introduction of methods such as Shifted Patch Tokenisation (SPT) and Locality Self-Attention (LSA).

Because of the limited dimensions of the dataset used, the approach employed during the training was that of **transfer learning**, a concept originating from educational psychology and adapted to computer science with the advent of large scale pre-trained models: just as human beings do when they generalise their experience to transfer knowledge across domains, with transfer learning ML models can approach their tasks by adapting the knowledge obtained by another model trained on a large dataset in a close domain [Zhuang et al., 2021]. Transfer learning is widely used in Computer Vision, with the models being pre-trained on the ImageNet database [Deng et al., 2009], consisting of more than 14 million images hand-annotated with information about the depicted objects, and then used as backbone of fine-tuned models trained on smaller amounts of task-specific data [Han et al., 2021].

As mentioned before, the complexity of the theoretical definition of the task forced the simplification of the problem from a multi-class classification to a **binary classification**: while with multi-class classification there is no opposition between two class labels, but rather a range of mutually exclusive specified classes [Sarker, 2021], with binary classification there are only two class labels (usually True/False). The training was performed on several different configurations of the model: the whole pre-trained model,

without freezing any weights; and only the classifier head with the frozen backbone pre-trained model, first with a shallow classifier (the one added automatically by the Hugging Face framework when loading the ResNet for Image Classification), then with a more complex structure with three fully-connected layers. These approaches shared the same first phase (loading and preprocessing the dataset, data collation and setup of the evaluation metrics), but differ in the second phase (loading the model, eventually freezing the weights and specifying a custom loss function). The best performing configuration was then fine-tuned using both the train and the validation set, and the resulting model is deployed via the test set to give a judgement on the final performance. The results and most relevant inaccuracies are then further discussed in consideration of the theoretical framework of the research.

Chapter 5

Results

Annotation process

The data annotation process was not straightforward from the outset. The widespread presence of colonial oppression in daily life necessitated a compromise between easily recognisable race-based triggers of sensitive content (e.g., graphic images, clear depictions of a race-based hierarchical social structure) and the numerous subtle signs of ongoing oppression (such as the display of medals and flags associated with the oppressor, as well as portraits of rulers). The taxonomy produced during the continuous discussions on the available data was used as an aid during the annotation process, but there were still some instances where its application was doubtful due to the issues and shortcomings of the definition of "sensitive content" presented in the previous chapter: the subjectivity of sensitivity, in fact, produced possibly unclear classes' boundaries especially regarding the class of "dubious content".

It goes without saying that this problem is particularly evident in the threefold definition, while in a binary definition the gap between the classes of "sensitive" and "not-sensitive" images is larger and the two classes are more easily recognisable as discrete. An example of a problematic annotation is shown in Figure 5.1: (a) was labeled as "dubious", while (b) was labeled as "sensitive". The difference in annotation is due to the perception of agency emerging from the pictures. While the labourers in picture (a) seem more relaxed and more self-assured, those in picture (b) give the impression of having been depicted in their role as employees, presumably lower in status than the photographer. However, this tendency to subjectivity is one of the clearest shortcomings of the annotation approach and should definitely be addressed and avoided.

This highlights the weakness of the threefold definition that, albeit being a reasonable attempt at representing the layered oppression of colonial regimes, does lack the necessary discrete structure necessary for an easier and smoother application to a Machine Learning pipeline. For this reason, the proposed definition of sensitive content has been simplified to a **binary definition**, distinguishing between "sensitive" and "not-sensitive" content, with the first category comprising of the original "sensitive" and "dubious" classes. The threefold definition will be used during the error analysis to assess the possible origin of the errors in the predictions generated by the validated model.

After the dataset preparation and annotation, different models were tested with different hyperparameter setups in order to explore the feasibility of addressing the research problem through an algorithm of image classification. In the end, the overall best performing model overall was deployed by validating its performance outside of the training process: after selecting the best performing configuration, the model was trained on the combination of train and validation set, and evaluated on the test set, which was never used during the development stage. The predictions were then analysed in correlation with the theoretical apparatus prepared beforehand and consisting in the three facets definition of sensitive content in an attempt to explain the possible cause of the errors.

Training setup

Due to its dimensions, the dataset was not entirely downloaded but rather used in **streaming mode** to avoid disk memory issues. This means that only small fractions of samples are loaded in memory and, as a consequence, the training is performed over a number of "steps" rather than "epochs". These two concepts are related: while an epoch is a complete cycle through the entire train set, a step is one optimisation update, with which the model parameters are adjusted based on a batch of the training data. The total number of steps performed during a whole training, therefore, depends on the number of epochs:

$$maxsteps = (trainsamples/batchsize) * numberepochs$$

Figure 5.1: Example of problematic annotations

(a) Leiden University Library, KITLV 68829 (b) Leiden University Library, KITLV 68880

Using as example this project's framework, with a batch size of 32, a train dataset comprising of a total of 1772 samples, and 15 training epochs, we can calculate the number of total steps as follows:

$$(1772/32) * 15 = 830.625$$

Because of this influence over the step parameter and its popularity in ML practice, the term "epoch" is preferred in the analysis of the different configurations and their results.

The choice of the **evaluation metrics** used to assess the performance of the models was also influenced by the dataset status. Despite accuracy being one of the most common evaluation metric for classification tasks, when used to assess the performance of a model trained on an imbalanced dataset as in our case it does not mirror the exact performance of the algorithm, but rather gives a misleading representation of it: indeed, accuracy is simply the proportion of correct predictions over the total number of predictions. However, in imbalanced classification problems, where the distribution of samples across the classes is not equal, accuracy can lead to an unreliable performance evaluation known as "accuracy paradox": a high score simply might be due to the heavy influence of the majority class, the samples of which are always correctly detected by a classifier which simply tries to maximise the accuracy [Galar et al., 2012].

For this reason, while all most common metrics for image classification tasks are measured to portray the performance of the models as accurately as possible, the one chosen to determine the overall best performing model is the **F1-score**, specifically measured in **macro average**. The F1-score is the harmonic mean of precision and recall thus merging these two competing metrics in one single measure. While precision is the proportion of correct predictions over the total number of predictions, recall is the fraction of correct predictions over the total number of samples that should have been predicted. These two metrics offer a trade-off: maximising precision reduces recall, and vice versa. As encompassing both, maximising F1-score is a good way to try to obtain the best possible classifier. Moreover, given the dataset imbalance between the classes, the use of the macro average helps in limiting the influence of the majority class on the metric score: while macro average computes each metric for each class independently and then measures the average, *de facto* treating equally all the classes, the micro average aggregates the contributions of all classes, giving equal importance to each data sample. For completeness, even the micro average for each metric is calculated, as it gives additional depth to the understanding of the model's performance.

Another popular measure used to assess the performance of a classification model is the **confusion matrix**: this kind of table can be used for both binary and multi-class classification and allows the

visualization of the predictions computed by the model, representing counts from predicted and actual values. This helps in understanding which classes are being incorrectly detected, and with which other class they are being confused. Figure 5.2 shows a 2x2 confusion matrix: the green main diagonal represents the correct predictions (True Positive, TP, and True Negative, TN), while the antidiagonal represents the incorrect predictions (False Positive, FP, and False Negative, FN).

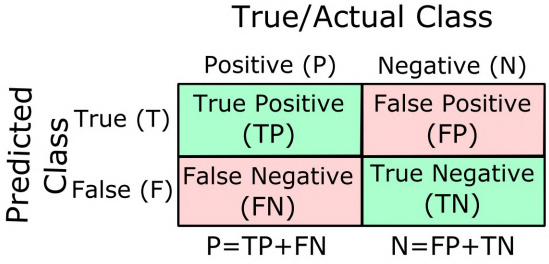

Figure 5.2: Illustrative example of a 2x2 confusion matrix. TP are the correctly classified positive samples, FN are incorrectly classified positive samples, TN are correctly classified negative samples and FP are incorrectly classified negative samples. From [Tharwat, 2020].

The experiments were all carried out using the Hugging Face platform and libraries. Within this framework, loading the model for image classification will automatically instantiate the selected model class with an image classification head on top (a linear layer on top of the pooled features). The specific checkpoint used in this research is "microsoft/resnet-50", which corresponds to v1.5 of ResNet, pretrained on ImageNet-1k at a resolution of 224x224. As previously mentioned, all of the experiments are accessible for further examination on a public interactive Weights & Biases dashboard, accessible from the Github public repository dedicated to this project [Borrini, 2024].

5.1 Binary classification

The binary classification approach was tested on three different ResNet architectures: in two of them the base model's weights are frozen and only the classifier head is trained, first in its base shallow form automatically provided by Hugging Face, then in a deeper structure consisting of three fully-connected layers with non-linear activation functions; finally, a third architecture is the whole model, which is finetuned without freezing any weights. These three architectures share some common parameters:

1. Batch size: 32.
2. Number of training epochs: 15 (as steps: 830.625, then increased for the most promising configurations).
3. Early stopping monitoring the evaluation loss score with patience of 5 evaluation calls (performed every 100 steps).

In this framework, the binary classes are encoded as follows: class "not-sensitive" is encoded with identifier "0", whereas class "sensitive" is encoded with identifier "1".

After the first experiments with different hyperparameter configurations on the base classifier head (baseC), only the best performing setups are also tested with the deep classifier head (deepC) and with the whole model (fullNet). Moreover, the same configurations for each architecture are tested out in two ways: first, by using the standard loss function of the model (which is the cross-entropy loss function in a `ResNetForImageClassification` model); then, by employing a custom weighted loss function using the normalized Inverse of Number of Samples (INS) weighting scheme, often adopted for imbalanced datasets [Huang et al., 2016, Wang et al., 2017] (baseC-w, deepC-w, fullNet-w). The weighted loss function will assign a higher penalty to errors made in the minority class, making the model more sensitive to this class by increasing its cost of misclassification.
The setup is shown in table 5.1:

5.1.1 Base classifier head

The base classifier head is the first architecture used to experiment the different hyperparameter configurations: it consists in the frozen ResNet50 base model with a shallow classifier head automatically added

	baseC1*	baseC1-w*	baseC2	baseC2-w	baseC3	baseC3-w
Learning rate	1e-2	1e-2	1e-4	1e-4	1e-6	1e-6
Weight decay	1e-3	1e-3	1e-5	1e-5	1e-7	1e-7
Weighted classes	No	Yes	No	Yes	No	Yes
	deepC fullNet	**deepC-w fullNet-w**				

Table 5.1: The configurations used for the binary classification runs. * indicates the best performing runs which have also been trained for a larger number of epochs

on top by Hugging Face during the loading of the model. It has the following structure:

```
Sequential(
    (0): Flatten(start_dim=1,end_dim=-1)
    (1): Linear(in_fetures=2048, out_features=3, bias=True)
)
```

The configurations tested with this architecture are shown in Table 5.2.

	baseC1	baseC1-w	baseC2	baseC2-w	baseC3	baseC3-w
Learning rate	1e-2	1e-2	1e-4	1e-4	1e-6	1e-6
Weight decay	1e-3	1e-3	1e-5	1e-5	1e-7	1e-7
Weighted classes	No	Yes	No	Yes	No	Yes

Table 5.2: Training runs for binary classification on shallow classifier

To understand the overall performance of each of these six models, we can compare the confusion matrices, shown in figs. 5.3 to 5.5. The only runs which seem to perform better than the others are runs **baseC1** and **baseC1-w**, although the custom weighted loss function does not seem to make a noticeable difference (the only difference between run baseC1 and baseC1-w is one prediction, which could also be due a normal random fluctuation of the model's performance.

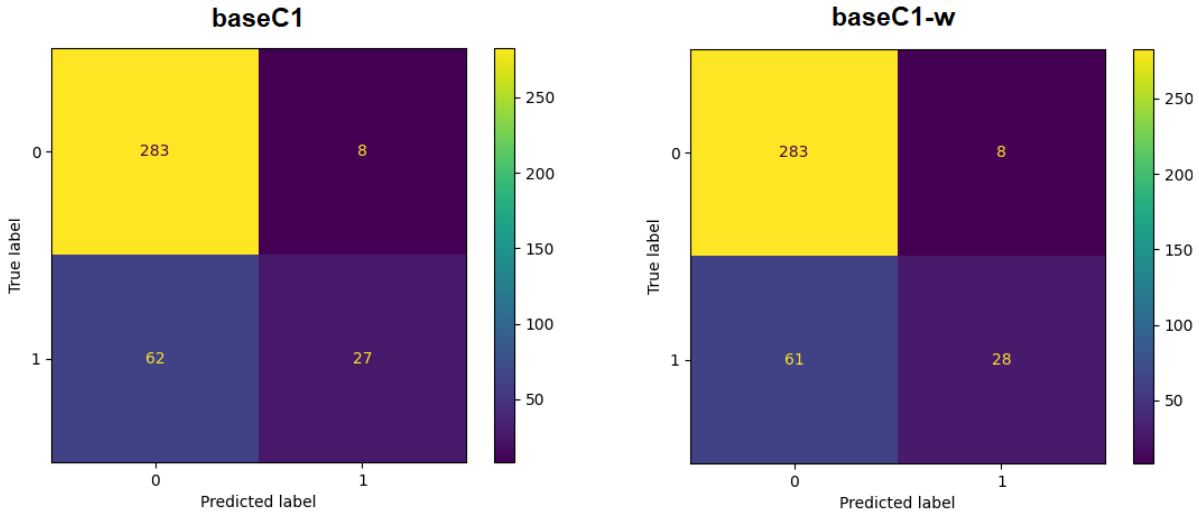

Figure 5.3: Confusion matrices of runs baseC1 and baseC1-w

The other models, shown in figs. 5.4 and 5.5, definitely confuse the two classes with one another, although in different ways:

- In runs baseC2 and baseC2-w, shown in 5.4, the algorithm detects only class 0 ("not-sensitive content") incorrectly classifying all of the samples belonging to class 1 ("sensitive content"). The

presence of a weighted custom loss function does not impact this result whatsoever, which is then probably caused by a suboptimal hyperparameter configuration.

- In runs baseC3 and baseC3-w, shown in 5.5, the algorithm predicts both classes with a preference for class 1 ("sensitive content"), the more scarce class, even when the prediction should be class 0.

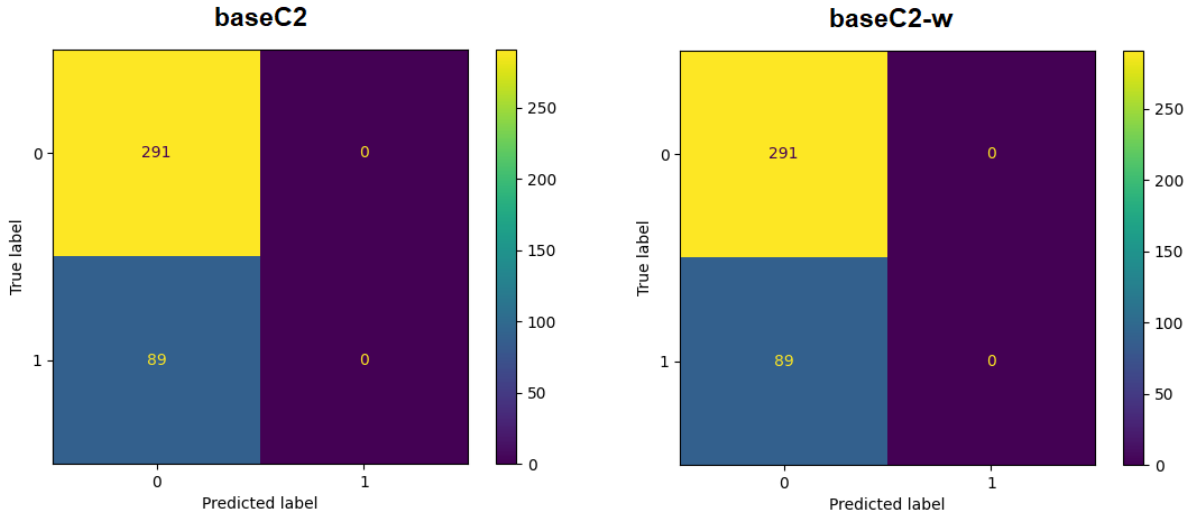

Figure 5.4: Confusion matrices of runs baseC2 and baseC2-w

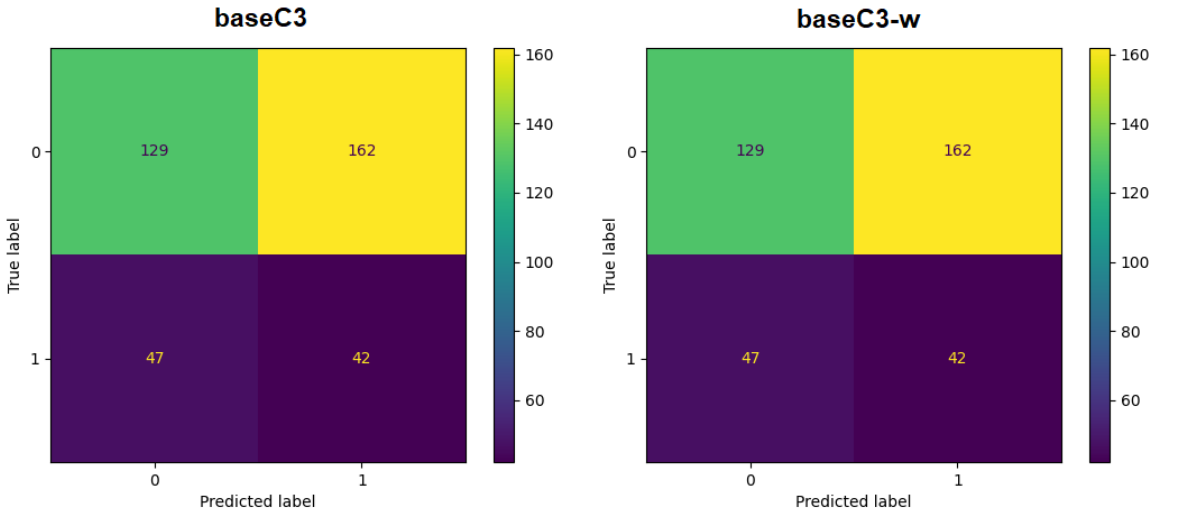

Figure 5.5: Confusion matrices of runs baseC3 and baseC3-w

Before moving onto the deep classifier and the full network architectures, two more tests with a double amount of training epochs are run with the best performing configurations of the base shallow classifier head, as shown in Table 5.3.

Again, the performance is evaluated on the confusion matrices of Figure 5.6, on which

By observing the confusion matrices in Figure 5.6 and comparing them to Figure 5.3, the decisive improvement is evident. Both matrices, in fact, tend to the ideal confusion matrix with all true positives (TP) and true negatives (TN) along the main diagonal, and zero false positives (FP) and false negatives (FN) in the off-diagonal entries. The result is still not ideal, with many wrong predictions for both classes, but this situation could be improved by further increasing the number of training epochs/steps.

	baseC1	baseC1-w	baseC1_30e	baseC1-w_30e
Number of training epochs	15	15	30	30
Number of steps	830.625	830.625	1661.25	1661.25
Learning rate	1e-2	1e-2	1e-2	1e-2
Weight decay	1e-3	1e-3	1e-3	1e-3
Weighted classes	No	Yes	No	Yes

Table 5.3: Further testing on runs baseC1 and baseC1-w

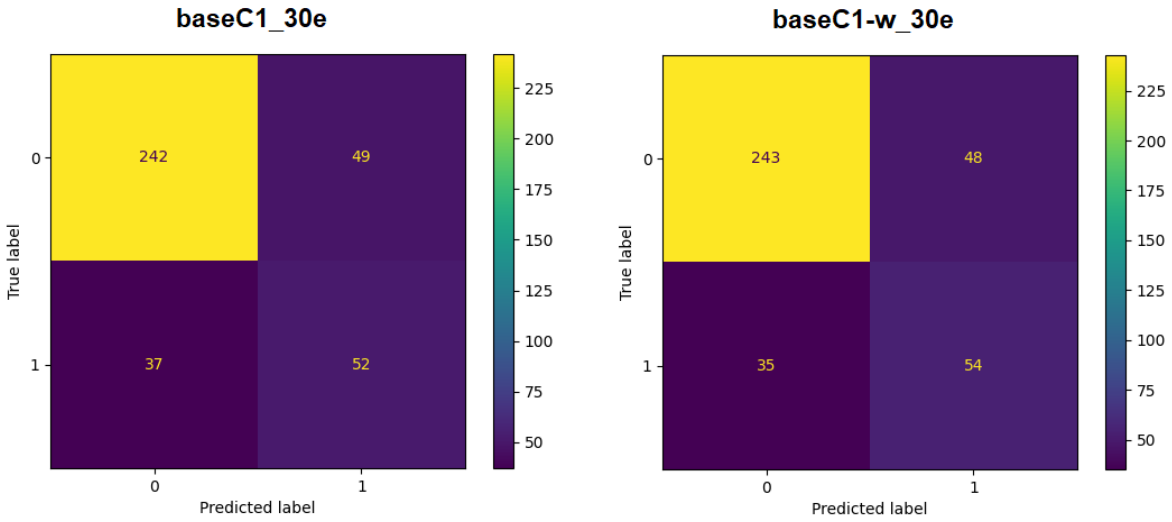

Figure 5.6: Confusion matrices of runs baseC1 and baseC1-w trained for 30 epochs

5.1.2 Deep classifier head

The second architecture employed is the deep classifier head, which is created by adding two more fully-connected layers with non-linear activation functions to the base classifier head, for a total of three fully-connected layers with non-linear activation functions.

Activation functions transform the summed weighted input from a node into an output value, either feeding it to the next layer, thereby activating the node, or as an output. In particular, the importance of non linear activation functions derives from their role in introducing non-linearity to the neural network, enabling the learning of more complex tasks through backpropagation. Among the several activation function, the one chosen for this project is Rectified Linear Unit (ReLu): this highly computationally efficient non-linear activation function was introduced in 2019 by [Agarap, 2019] and has been mostly used in CNNs, solving the problem of gradient disappearance caused by the increase of network layers. Although ReLu has since been improved with the introduction of minor modifications, because of the experimental nature of this research we have decided to use its earliest definition. This decision was made to avoid introducing another variable to the problem.

The structure of this deep classifier is the following:

```
Sequential(
    (0): Flatten(start_dim=1, end_dim=-1)
    (1): Linear(in_features=2048, out_features=1024, bias=True)
    (2): ReLU()
    (3): BatchNorm1d(1024, eps=1e-05, momentum=0.1, affine=True,
           track_running_stats=True)
    (4): Linear(in_features=1024, out_features=512, bias=True)
    (5): ReLU()
    (6): BatchNorm1d(512, eps=1e-05, momentum=0.1, affine=True,
           track_running_stats=True)
    (7): Linear(in_features=512, out_features=2, bias=True)
)
```

As mentioned, not all previous configurations were run in this architecture. Only the best performing

runs were replicated, namely the ones shown in table 5.4:

	deepC	deepC-w
Learning rate	1e-2	1e-2
Weight decay	1e-3	1e-3
Weighted classes	No	Yes

Table 5.4: Training runs for binary classification on deep classifier

However, from the inspection of the confusion matrix for these two runs, in Figure 5.7, we can see how increasing the depth of the classifier did not entail an improvement of the results: in fact, the matrices show a lot of indecision in the distinction between the two classes, incurring in the same kind of error of models baseC2 and baseC2-w.

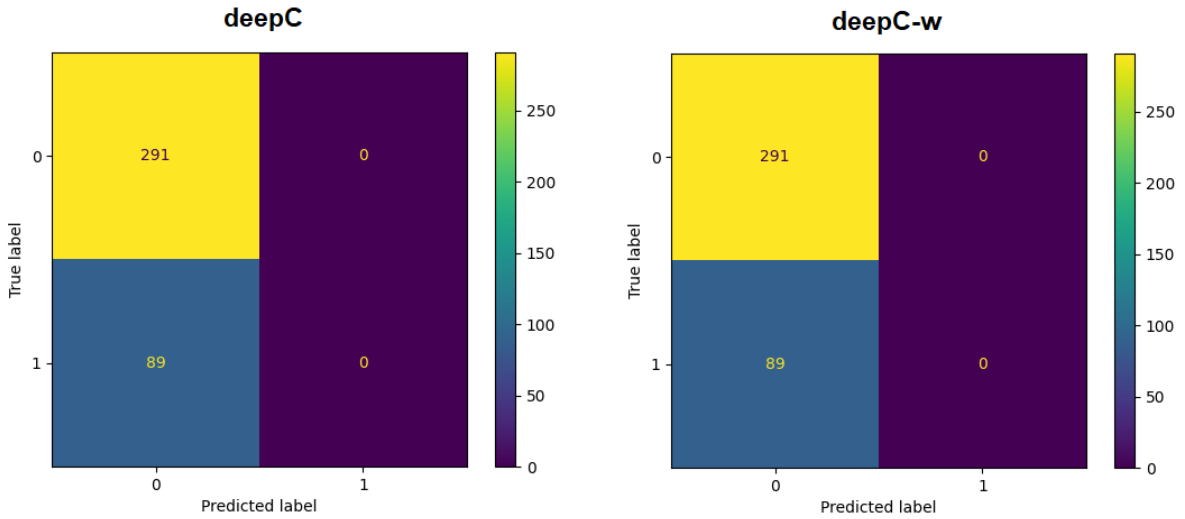

Figure 5.7: Confusion matrices of runs deepC and deepC-w

5.1.3 Finetuning

Eventually, the training was performed on the whole architecture, without freezing any weights. Again, the structure of the architecture is that automatically provided by Hugging Face during the loading of the model: attached to the ResNet module, there is an additional simple classifier head. The training, just as with the deep classifier head, was performed only with the best performing hyperparameter configurations from the runs with the simple shallow classifier, as shown in table 5.5.

	fullNet	fullNet-w
Learning rate	1e-2	1e-2
Weight decay	1e-3	1e-3
Weighted classes	No	Yes

Table 5.5: Training runs for binary classification on the full network

Nonetheless, even in this configuration the performance was subpar, probably due to the high amount of parameters to train. Once again, the low performance is confirmed by the confusion matrices in Figure 5.8, which again resembles the matrices of runs baseC2, baseC2-w, deepC and deepC-w.

5.1.4 Validation

Finally, all of the runs were evaluated through the most common evaluation metrics for an image classification task: accuracy, precision, recall, and F1-score. Among these, the latter three were measured using both micro and macro average to give an overall view of the performance of each model as complete as

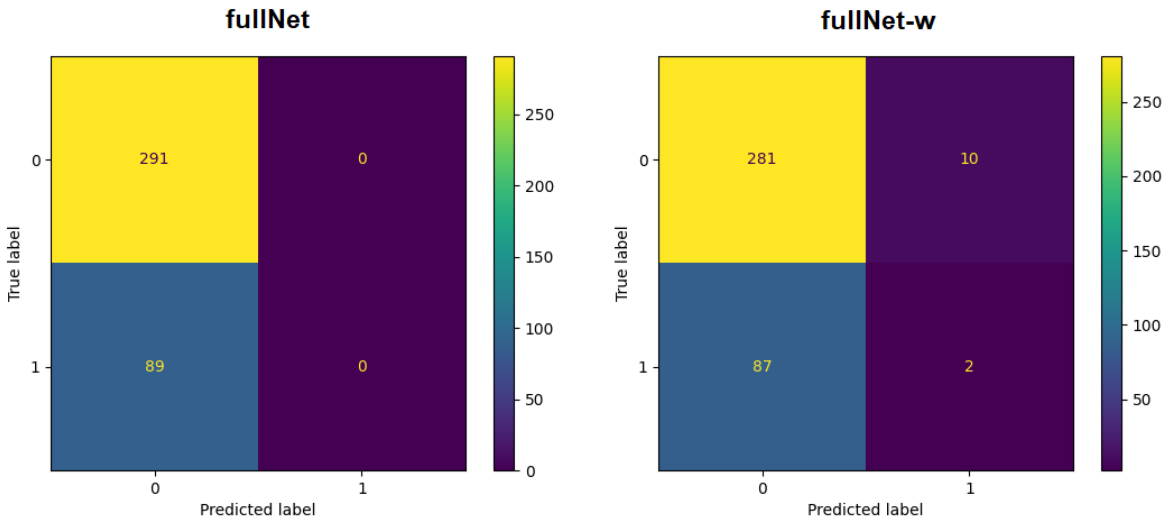

Figure 5.8: Confusion matrices of runs fullNet and fullNet-w

possible. The results are presented in Table 5.6. The metric used to assess the best model was F1-score in macro average. As predictable from the inspection of the confusion matrices, the highest F1-scores were calculated for the runs baseC1 and baseC1-w, the two runs in which the training was performed merely on the base shallow classifier head with a learning rate of 1e-2 and a weight decay of 1e-3. It is worth noting how the application of a custom weighted loss function had no effect on the performance of the model for all configurations except for baseC1, baseC1-w, fullNet and fullNet-w, in which it entailed a slight improvement.

Run ID	**Accuracy**	**Precision**		**Recall**		**F1-score**	
		Micro	Macro	Micro	Macro	Micro	Macro
baseC1	**0.815**	**0.815**	**0.795**	**0.815**	**0.637**	**0.815**	**0.662**
baseC1-w	**0.818**	**0.818**	**0.800**	**0.818**	**0.643**	**0.818**	**0.669**
baseC2	0.765	0.765	0.382	0.765	0.5	0.765	0.433
baseC2-w	0.765	0.765	0.382	0.765	0.5	0.765	0.433
baseC3	0.45	0.45	0.469	0.45	0.457	0.45	0.419
baseC3-w	0.45	0.45	0.469	0.45	0.457	0.45	0.419
deepC	0.765	0.765	0.382	0.765	0.5	0.765	0.433
deepC-w	0.765	0.765	0.382	0.765	0.5	0.765	0.433
fullNet	0.765	0.765	0.382	0.765	0.5	0.765	0.433
fullNet-w	0.744	0.744	0.465	0.744	0.494	0.744	0.446

Table 5.6: Evaluation metrics scores for the binary classification runs

Consequently, the configuration chosen to perform a validation of the model was that of run **baseC1-w**. The validation training run was performed on an amplified dataset: both the train and the validation set were used as input training data, and the model's performance was evaluated on the test set. Due to the technical and time restrictions of this project, it was possible to perform deployment of the model only through a training session of a total of 40 epochs as shown in Table 5.7. Still, the performance proved an improvement towards the ideal confusion matrix, as seen in Figure 5.9.

Moreover, the progressive improvement of the confusion matrix's results in the runs sharing the same hyperparameter configuration and differing only for the number of training epochs (and the dataset, which for run valid_baseC1-w is enhanced) is optimistic for further improvement on a larger number of training epochs. This is illustrated in Figure 5.10.

	baseC1-w	baseC1-w_30e	valid_baseC1-w
Number of training epochs	15	30	40
Number of steps	830.625	1661.25	2215
Learning rate	1e-2	1e-2	1e-2
Weight decay	1e-3	1e-3	1e-3
Weighted classes	Yes	Yes	Yes

Table 5.7: Configuration of the validated model compared to previous best performing runs baseC1-w, baseC1-w_30e.

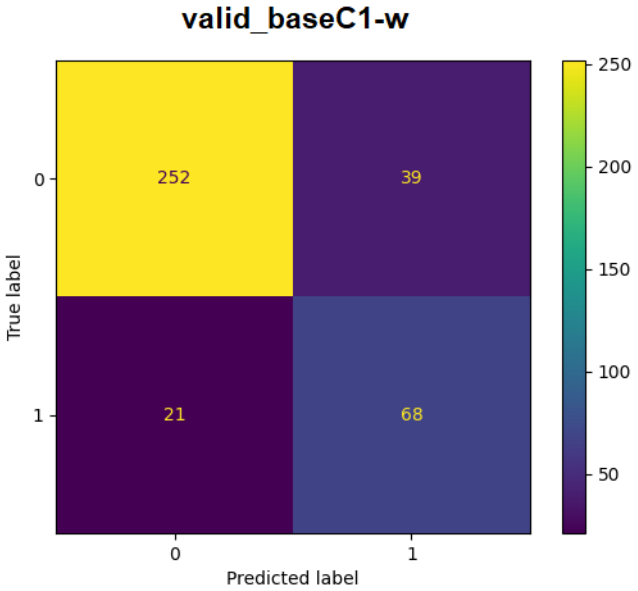

Figure 5.9: Confusion matrix of validated model

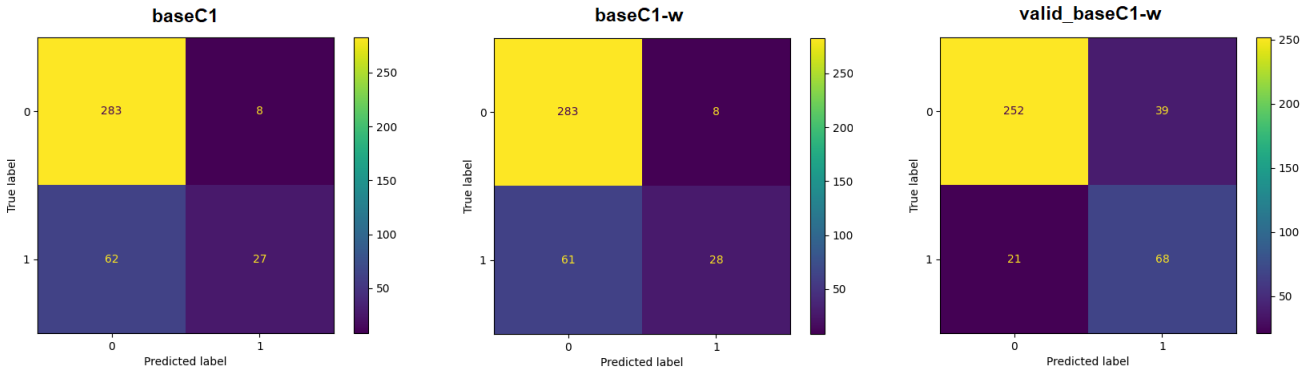

Figure 5.10: Confusion matrix of validated model compared to previous best performing runs baseC1-w, baseC1-w_30e

Finally, the improvement in performance is also confirmed by the metrics, shown in Table 5.8, where the best performing model is emphasised in bold.

Run ID	Accuracy	Precision		Recall		F1 score	
		Micro	Macro	Micro	Macro	Micro	Macro
baseC1-w	0.818	0.818	0.800	0.818	0.643	0.818	0.669
baseC1-w_30e	0.781	0.781	0.701	0.781	0.720	0.781	0.709
valid_baseC1-w	**0.842**	**0.842**	**0.779**	**0.842**	**0.815**	**0.842**	**0.793**

Table 5.8: Evaluation metrics scores for the validated model

5.2 Error analysis

To gain a more thorough understanding of the performance of the validated model, the incorrect predictions were selected and analysed in order to understand the possible reasons behind the confusion between the two classes. The approach used was a continuous comparison between the incorrect predictions, the annotated data of the train set, and the definitions for the categories of the two classes developed in Chapter 3. Referring to the confusion matrix of the validated model presented in Figure 5.9, the incorrect predictions are the ones in the antidiagonal of the matrix:

- **False "sensitive"**: in the top-right corner of the confusion matrix, it is the images which have been mistakenly detected as "sensitive", while their true label would be "not-sensitive"
- **False "not-sensitive"**: in the bottom-left corner of the confusion matrix, it is the images which have been mistakenly detected as "not-sensitive", while their true label would be "sensitive"

The situation is summarised in Table 5.9.

	Number of images	Figure
False "sensitive"	38*	5.11
False "not-sensitive"	21	5.18

Table 5.9: The distribution of errors in the predictions of the validated model. The number of false "sensitive" predictions is one sample less than in the confusion matrix of Figure 5.9, due to the inherent stochastic nature of neural networks.

Furthermore, it is important to note once again the distinction between the "sensitive" and the "not-sensitive" classes in the binary classification (with the previously "dubious" category definition being denoted by the + symbol):

- **Sensitive content**: content which is more explicit and easily recognisable as immediately sensitive or which could cause offence to Indigenous communities if not presented with additional consideration.
  - Any document reiterating the colonial racist, xenophobic, and discriminatory beliefs, providing harmful and stereotypical representation of Indigenous individuals and communities and demonstrating, in general, bias and exclusion of marginalized people.
  - Any document depicting explicit violent graphic content including death, medical procedures, crimes against the person, war, and terrorist acts.
  - Any document containing clear symbols and references to the colonial context (e.g., foreign invaders' flags, portraits of the rulers, medals...).
  + Any document that could be triggering or cause offence to the Indigenous communities affected by colonial imperialism if not presented with the due consideration and empathy (e.g., studio portraits, Cultural Heritage with people...).
- **Not-sensitive content**: content which does not display any clear or explicitly sensitive feature (e.g., crowd pictures, landscapes, catalog pictures of Cultural Heritage and common objects...).

False "sensitive"

The analysis starts with the inaccurate false "sensitive" predictions, as shown in Figure 5.11, and is carried out by detecting some recognisable types of photography.

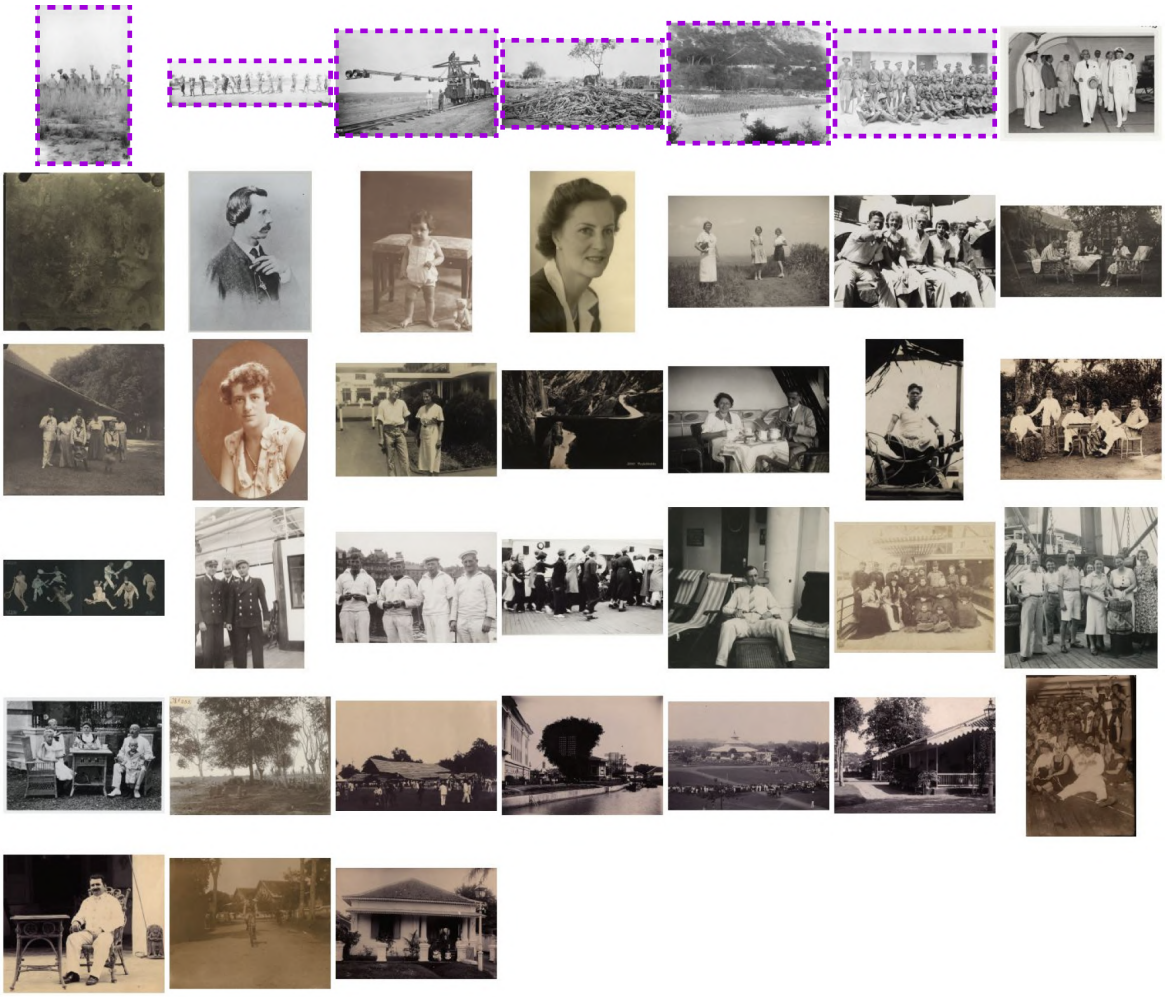

Figure 5.11: False “sensitive” predictions: “not-sensitive” images mistakenly detected as “sensitive”. The images from the IWM collection are highlighted with a purple dotted line and will not be reproduced on a larger scale due to copyright restrictions.

Upon initial inspection of this batch, the presence of **studio portraiture photography** is evident: this type of image has very defined characteristics, with high-contrast between two well separate areas, usually a blank or plain background and the posing subject, and it originally pertained the "dubious" category, now full-fledged "sensitive content". For some of the samples presented in Figure 5.12, it is clear how the model corrected errors occurred during the annotation process. However, it also inaccurately predicted as "sensitive" other images which do share some visual features with more typical studio portraiture photography.

Figure 5.12: False "sensitive" predictions - Studio portraiture. Most of these pictures are clear examples of studio portraiture and their erroneous annotation was corrected by the model. Therefore, only 3 pictures, here highlighted, are incorrectly predicted: albeit similar to studio portraiture, they were annotated as "not-sensitive" for different reasons. The ones in red due to their more natural and spontaneous poses, while the one in orange is a photomontage. Examples from the train set are not provided for the clarity of the studio portraiture concept.

Another recognisable group of pictures is that of **landscape pictures**, shown in 5.13, which are a common instance of the "not-sensitive" class. Because of their distinctive features, landscape pictures do not run the high risk of being confused with something else and it would not be an overstatement to speculate that these errors could be due to the low number of training epochs, still only 40 for the validated model. The constant improvement in performance of the configuration baseC1-w (as detected in Figure 5.10, where the confusion matrices for runs baseC1-w, baseC1-w_30e and valid_baseC1-w are compared) contributes to this optimistic hypothesis.

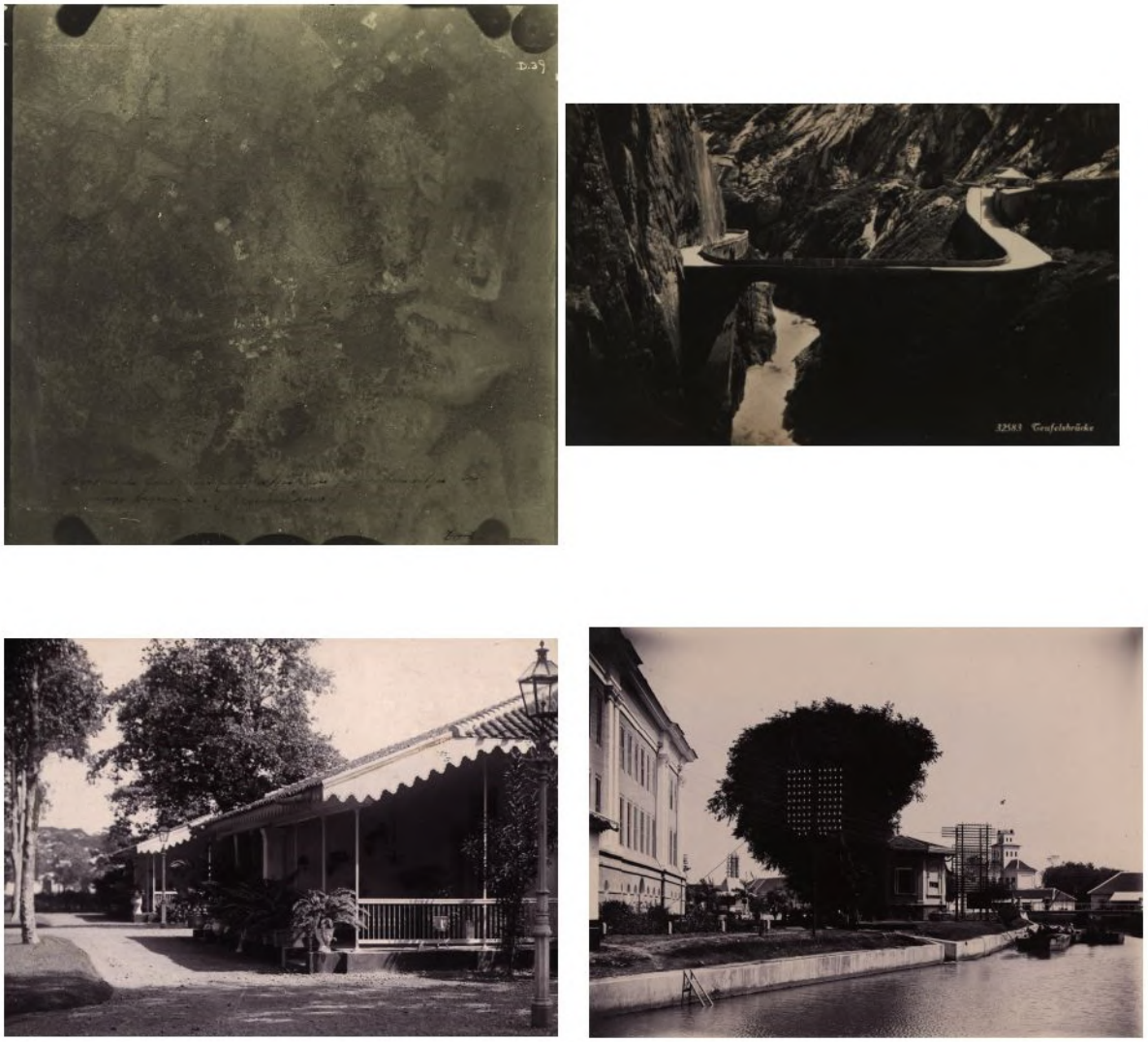

Figure 5.13: False "sensitive" predictions - Landscapes

Populated landscapes differ slightly from landscape photography in that they include people (or crowds), which is a feature affecting the annotation. The incorrect predictions are displayed in Figure 5.14 and they could be due to the presence of similar images in the train set which are labeled as "sensitive" for the presence of specific details that more explicitly denote a connection to the colonial oppression.

Figure 5.14: False "sensitive" predictions - Populated landscapes. The images highlighted in blue are examples from the train set, added to highlight the similarities that could have caused mistakes in the prediction. The details which triggered the annotation are shown in the zoom-in. Under this category also falls a picture of the IWM collection depicting troop deployment, not reproduced for copyright restrictions.

A separate case is that of **cultural heritage pictures**, shown in 5.15. The annotation for this type of image highly depended on specific details regarding the setting and the entities depicted: if the picture only depicted a cultural object and was taken in a studio, the annotation was "not-sensitive"; however, when the setting was a natural one, it heavily relied on the presence of people, who determined a possible stereotypical representation and, therefore, the "sensitive" label. However, the framing profoundly influences the detection of these details, therefore affecting the class prediction.

Figure 5.15: False "sensitive" predictions - Cultural Heritage. The images highlighted in blue are examples from the train set, shown to highlight the similarities that could have caused the error in the prediction. The details which triggered the annotation are shown in the zoom-in.

The last recognisable type of photography is that of **lifestyle documentary**. Even in this case, it is very probable that the cause of the error was the difficulty of the model to detect the details triggering the annotation as "sensitive", and rather relying on larger features of the picture to label it such as, for example, the presence of large buildings and groups of people. The situation is exemplified in Figure 5.16.

Figure 5.16: False "sensitive" predictions - Lifestyle documentary. The images highlighted in blue are examples from the train set, shown to highlight the similarities that could have caused the error in the prediction. The details which triggered the annotation are shown in the zoom-in.

For what concerns the remaining pictures, shown in Figure 5.17, it is hard to define exactly possible error categories: while being fairly similar to lifestyle documentary photographs, they are not part of the same genre and rarely share the same type of large visual feature. Once again, the incorrect predictions could be due to the presences of "sensitive" images with similar features in the train set which might have been distinguishable only by small details, not detected by the model, which then learned to use larger features such as the subjects distribution, the distribution between ground and sky, and the presence of buildings.

Figure 5.17: False "sensitive" predictions - Others. Under this category also fall five picture of the IWM collection, not reproduced for copyright restrictions.

Table 5.10 summarises the overall situation of the error analysis on false "sensitive" predictions.

Misclassified photo style	Number of images	Cause of misclassification	Figure
Studio portraiture	**8***	**Errors and similarity**	**5.12**
Populated landscapes	4	Emphasis on large features	5.14
Landscapes	4	Few training epochs	5.13
Cultural Heritage	**1**	**Emphasis on large features**	**5.15**
Lifestyle documentary	2	Emphasis on large features	5.16
Other	21	/	5.17

Table 5.10: The different types of errors in the misclassifications of false "sensitive". In bold the styles belonging to the sub-category of "dubious content". For what concerns studio portraiture, 5 of the 8 pictures were, in fact, correctly predicted as "sensitive", *de facto* making up for mistakes in the annotation, and only 3 pictures were actually incorrect predictions.

False "not-sensitive"

The second part of the error analysis regards the false "not-sensitive" predictions, displayed in Figure 5.18. These are "sensitive" pictures which were incorrectly predicted as "not-sensitive". To examine this situation we can use the same method employed for the false "sensitive" predictions, looking for familiar overarching types of photography and possibly using the ones already introduced in the first part of the error analysis.

Figure 5.18: False "not-sensitive" predictions: images "sensitive" mistakenly detected as "not-sensitive". The pictures from the IWM collection are highlighted with a purple dotted line and will not be reproduced on a bigger scale for copyright reasons.

From a first glance we can see that even in this batch of pictures there are **studio portraits** (Figure 5.19). This genre is not only strictly connected to the "sensitive" class, as there are many examples of it throughout the whole dataset, but this relation can also be reinforced by the presence of additional triggers for the "sensitive" label such as medals. However, as we have seen in the first part of the error analysis, it is unlikely that the model is detecting this sort of smaller features. Still, given the exact and

precise definition of studio portraiture, the presence of similar pictures in the "not-sensitive" class which could have negatively influenced these predictions is highly unlikely, except for the presence of errors in the annotation phase. Therefore, these incorrect predictions might be resolved by simply increasing the number of training epochs/steps.

Figure 5.19: False "not-sensitive" predictions - Studio portraiture. The mistakes in the predictions could be due, in this case, to a low number of training epochs, as this type of photography is very well defined and has striking features which would make it hard to incorrectly detect.

Crowd and populated landscape photographs' labeling heavily relies on the feasibility of detecting smaller features such as, for example, race-based hierarchical structures (described as "structural discrimination" in the taxonomy) in wide shot photography. However, given the previous steps of the error analysis, it is possible that this kind of small features was not detected by the model and, therefore, the pictures were considered annotated based on larger features normally associated with the "not-sensitive" class. This is the case for most of the pictures from the IWM collection, which will not be reproduced in larger scale for copyright reasons, but other examples are shown in Figure 5.20.

Figure 5.20: False "not-sensitive" predictions - Populated landscapes

Finally, the last type of error in the false "not-sensitive" predictions regards **cultural heritage pictures**. The images of this category, originally belonging to the "dubious" sub-class, probably do not get recognised as "dubious" due to the lack of detection of the people present within them. Moreover, the high number of pictures of cultural objects in the "not-sensitive" class of the train set probably positively influenced their identification. The situation is depicted in Figure 5.21.

As in the first phase of the error analysis, even in this case there are some incorrectly predicted images for which is hard to speculate the cause (Figure 5.22). The probable reason behind these incorrect predictions is the presence of "not-sensitive" images with similar features in the train set, the inability of

Figure 5.21: False “not-sensitive” predictions - Cultural Heritage

the model to detect the detail triggers and its tendency to generalise on the overall content of the image.

Figure 5.22: False “not-sensitive” predictions - Others

Table 5.11 summarises the situation of the error analysis on false “not-sensitive” predictions.

In the end, as we can also see from Table 5.12, which displays the overall results of the error analysis, four different possible reasons for incorrect predictions by the model were found:

1. **Errors in the annotation phase**: it goes without saying that the presence of several errors in the annotation may be the first cause of incorrect predictions during the deployment of the model. This is usually prevented by preparing a pilot phase for the annotation, during which both the task definition and the annotation guidelines are refined little by little, ensuring an annotation process

Misclassified photo style	Number of images	Cause of misclassification	Figure
Studio portraiture	**3**	**Few training epochs**	**5.19**
Populated landscapes	3	Emphasis on large features	5.20
Cultural Heritage	**4**	**Emphasis on large features**	**5.21**
Other	6	/	5.22

Table 5.11: The different types of errors in the misclassifications of false "not-sensitive". In bold the styles belonging to the sub-category of "dubious content".

Incorrect prediction type	Photo style	Samples' number	Cause
False "sensitive"	Studio portraiture	5+3	Errors and similarity
	Populated landscapes	4	Emphasis on large features
	Landscapes	4	Few training epochs
	Cultural heritage	1	Emphasis on large features
	Lifestyle documentary	2	Emphasis on large features
	Others	21	/
False "not-sensitive"			
	Studio portraiture	3	Few training epochs
	Populated landscape	3	Emphasis on large features
	Cultural heritage	4	Emphasis on large features
	Others	6	/

Table 5.12: Overall final results of the error analysis.

as homogeneous and precise as possible.

2. **Few training epochs**: this is the case for those image types which have very clear and distinctive features such as studio portraiture and landscape. Possible erroneous predictions in these two types of images are due to either the natural stochastic nature of neural networks, which improve the performance and the learning of the mapping function as the number of training epochs or steps increase.

3. **Feature similarity**: as in the case of false "sensitive" predictions of studio portraiture photographies, it is possible that, regardless of the precision of the definition of a type of image, its visual features and characteristics may be shared across different genres and samples. This is particularly striking when the model cannot make use of the colour mode attribute of the images, which is a highly semantic feature and would be extremely helpful in distinguishing between samples: while with sepia or grayscale images both the sky and a plain background have the same colour and therefore, are not distinguishable between one another, with coloured images this is not the case. The presence of colour undoubtedly provides an additional factor of differentiation between similar images.

4. **Emphasis on larger features**: by analysing both kinds of errors made on populated landscapes images, a pattern seems to emerge. In fact, it looks like the model, during the training, probably puts more emphasis on different features depending on the class. While this surely influences *how* the model learns, provoking unstable prediction results for the same type of photography (again, the example would be that of populated landscapes), this is not the only consequence: in fact, the emphasised features seem to be always the large features of the images. As a matter of fact, all instances in which the triggering factor was on a small scale with respect to the overall image have faced some issues during the prediction, being assigned to the opposite class.

Chapter 6

Discussion

With this research, we attempted at testing the feasibility of addressing the issue of automatic sensitive content detection in colonial photographic archives through the employment of an Image Classification task and model. The liveliness of this research field is testified by the projects applying AI technologies and techniques to digitised archival collections mentioned in chapter 2. Specifically, this research positions itself as a small followup to the EyCon project [Aske and Giardinetti, 2023], attempting at experimenting with the use of Computer Vision techniques to address the identification of sensitive content. The importance given to this aspect of archival research is born from the postmodernist inquiry on the role of archives as institutions and from the stress on greater transparency and accountability, calling for a greater awareness of the diversity and ambiguity inherent to these information systems and their users [Cook, 2001]. This research fits the postmodernist call for emphasising the diversities of marginalised voices, and does so by employing modern technologies such as AI and neural networks.

The analysis of the research conducted identifies possible feasible research avenues to address the problem of automatic sensitive content detection in colonial photographic archives, albeit at the cost of simplifying the issue to keep it within the general scope and reach of an image classification task. Overall, established a correct, thorough and homogeneous annotation phase, the application of binary image classification algorithms for colonial sensitive images seems to be a **feasible application of ML** to archival institutions. The recognition of the different aspects of colonial oppression from the simple image content, though, is a harder task that would require either a change in approach (by employing, for example, Object Detection based Image Classification) or by making use of the archival metadata available, which certainly allows for a more thorough interpretation of the images at hand.

The error analysis identifies several possible issues emerging from the use of a very general Machine Learning task like Image Classification. In fact, larger projects such as the EyCon have employed different methods, like Object Identification, to attempt at addressing a similar issue. However, in this case we wanted to experiment with the overall interpretation of the image, trying to give a more rational and scientific view of the pervasive oppression carried out by colonial empires and trying to describe it in discernible elements. As such, it encounters entirely different problems, from the **possible fragility of the definition and annotation**, to the **inadequacy of the chosen task**.

In fact, the fragility and opacity of the classes' boundaries violently emerges from the results of the error analysis. This is particular evident in those cases in which the annotation was mainly determined by very small details in the images. These features proved to be too small to be detected by the trained model, which therefore learned an incorrect mapping function between the label and the wrong set of features, likely the larger ones, which are easier to detect. This problem, very common in image classification task, is also known as **feature scale variation** and it is very common in the dataset used in this research, as the annotation often depended on the detection of these minor features, sometimes difficult to detect even for the human annotators. Moreover, the biases in learning cause by the feature scale variation can also be affected by the imbalance present in the dataset classes. Some photography types which have specifically encountered this type of error are populated landscapes and cultural heritage photography. The choice of an Image Classification algorithm automatically entailed other notable problems such as **intra-class variation**, **similarities across classes** and **cluttered images**. The first two errors are further enhanced by the colour similarities between historical pictures, which variate between grayscale and sepia colour modes. Of course, this characteristic is highly disadvantageous for a clear differentiation of the environmental features of the image (e.g., a cloudless sky can easily be confused with a blank studio background), but also for the detection and differentiation between different small features, even if they were actually detected by the model (which, again, is not the case, as highlighted by the feature scale variation problem). Another common problem also encountered in the dataset used for this research

is that of cluttered environment: the presence of many different objects in scenes means a high level of visual complexity and that makes it difficult to discern specific details (again, sometimes fundamental for the annotation in our case), or extract meaningful information. In all effect, the analysis and interpretation of the images during the annotation phase was also made more challenging by these chaotic and crowded images. Dealing with cluttered images often requires specialized image processing techniques, such as image segmentation or feature extraction, to isolate and identify relevant visual elements within the complexity. These techniques have not been put in place during this project due to its general restrictions in terms of time, hardware, workforce. Other challenges regard the **dimensions, quality, and balance of the dataset**. In fact, a sufficient number of high quality training samples is a prerequisite for a successful classification and, unfortunately, the dataset used for this research comprised of not only very few images for a Computer Vision task (the more complex the problem at hands, the more images for each class will be needed to ensure an accurate learning of the mapping function), but also very few high quality samples: the only high quality files were the IWM images, which were provided directly by the institution and amounted to only 199 samples; the remaining 1978 files were webscraped, resulting in a lower quality. Moreover, as already stated, the dataset was also imbalanced, with 1939 samples of "not-sensitive content" and only 593 samples of "sensitive content" (respectively, 330 "dubious content" and only 263 "sensitive content"). Clearly, this disparity heavily affects the quality of the learning.

Regardless of these issues, there was an overall observable improvement of the performance of the binary classification model in the configuration of the base shallow classifier with a custom weighted loss function (id: baseC-w). Due to the time restrictions of this project, the validation of this configuration was done only over 40 training epochs, but the training and validation loss curves show **optimistic results**. Because of this, we think that the use of a binary classification model could be a good starting point for then further analysis of the "sensitive" class (encompassing both the original "sensitive" and "dubious" classes). Elucidating the distinction between "sensitive" and "not-sensitive" images, describing clearer borders between the two classes, could certainly bring more improvements to the algorithm. Then, the different layers of the colonial oppression could be detected by employing other Computer Vision algorithms such as Object Identification, which could help detect the small features that were lost using Image Classification. The object detection results could also be further improved by employing clothing style estimation or automatic colourisation of historical images to reach a clearer identification of the small features triggering the detection of sensitive content.

The development of this research encountered several limitations:

1. The **complexity of the task** would require further insight from a variety of possible stakeholders from both archival sciences, historians and diasporas communities to ensure a profound understanding of the problem is reached. Understanding and clearly defining sensitive content in the context of colonial archives was, from the beginning, an obviously very subjective task. Given the nature of such archives, filled with hurtful stories and histories, at first a threefold definition of "sensitive content" was defined, in an attempt at encompassing the breadth of such notion. However, due to the impossibility of discerning the different degrees and stages of the oppression put in place by colonial regimes from the simple observation of the visual features of the images, the definition was simplified to a binary definition, opposing the categories of "sensitive" and "not-sensitive", at the price of a shallower comprehension of its inherent complexity.

2. **Using only the visual contents** of an archival record means stripping it of possibly hundreds of years of layered historical information and nuance. The use of the archival metadata and cataloguing information would certainly refine the annotation choices, making it more precise and allowing a more comprehensive understanding of the complexities of colonial empires and the connected archival institutions. This was tested in the annotation phase, during which the need for continuous discussion on most data soon became clear: very often the assigned label would change when confronted with new information regarding the people portrayed in the pictures, the context surrounding the shooting of the picture, the photographer identity or even just the title given to the picture in the digital database. This highlights the importance of additional information for a better performance of the model and introduces the next limitation of the research.

3. It is highly possible that the **model and task** employed were **too general** to detect the specificities of colonial oppression: when the annotation of the picture depended on small details (and this is the case for many pictures in the "sensitive" category) the model did not detect those small triggers, but rather the whole structure of the picture, therefore annotating with the same label other images which shared the same big structure but lacked the small details. This means that, in most positive cases, the incorrect predictions are due to this failure of the model in detecting the smaller features, which frequently were considered as distinctive traits for the differentiation between the two classes.

The presented limitations and issues could be overcome in feature research, this project standing as a starting point for further in-depth analysis on the status of sensitive content in colonial photographic archives. Given the fuzzy definitions of sensitivity, a concept which remarkably does not have a solid and accepted common definition but rather changes and is adapted to the context of use, the analysis on the different facets and scopes of this notion in the context of colonial photographic archives surely can work as a starting point for further refinements and contributions, even when not used as a prerequisite for the application of Machine Learning methods to archival sciences, institutions and collections. Stripping it from this limiting scope, which clearly marks the necessity for a compromise between the depiction of the several intricate layers of sensitivity in the context of colonial enterprises and the ease of computation, surely would open up several possibilities in the development of a structured, layered, definition and taxonomy rendering the different degrees and dimensions of sensitivity in the context of hurtful cultural heritage.

For what concerns the technical aspect of this research, surely several improvements could be made while staying within the context of Image Classification. Some first features to improve, even for a simple binary classification, regard the number of annotators, their preparation, and the collection of a larger dataset to further improve the promising results. While in this research no additional image preprocessing was put in place, as usually digitised photographic archival records are standardised and share the same layout, some interesting processing which could be applied would be, for example, automatic image colourisation to make up for the large degree of noise in busy and crowded pictures. Then, different hyperparameter configurations should be tested to ensure enough experimenting on most, if not all, possible variables in the Machine Learning discourse. The best configuration in the course of this research surely could work as a starting point for further testing. Then again, other Computer Vision tasks are worth exploring, such as Object detection or, expanding the horizon to new and little explored research enterprises, multimodal learning, which integrates the information coming from different modalities exploring the interaction between various types of data. This approach is the one that would probably lead to the best results given the intrinsically multimodal nature of photographic archival records, which are made of not only the mere visual information, but also and importantly the additional textual metadata and archival descriptions, which allow the user to take a glimpse into the depth of the record. Some recent and popular models worth experimenting with could be Grounded Language-Image Pre-training (GLIP) [Li et al., 2021] and Contrastive Language-Image Pre-training (CLIP) [Radford et al., 2021].

Chapter 7

Conclusion

In conclusion, this research embarked on an exploration of the feasibility of utilizing image classification techniques to tackle the complex issue of automatic sensitive content detection within colonial photographic archives. It is evident from the extensive review of existing projects and technologies that the integration of advanced ML methodologies in digitised archival collections is gaining momentum, with endeavors such as the EyCon project exemplifying the resolution to foster critical engagement with colonial warfare history and paving the way for innovative approaches specifically regarding the potentially sensitive content of colonial archives. The significance of discerning this type of content stems from postmodernist inquiries into the nature of archives as sites of contested power, urging for heightened transparency and accountability in handling diverse and multiple historical narratives, particularly concerning marginalised communities, to prevent the risk of only relocating the problem of cultural Western hegemony and viewpoints in traditional archival practices. By leveraging modern technologies like AI and neural networks, this study aligns with the postmodernist imperative to amplify these silenced voices and shed light on the multifaceted dimensions of colonial oppression.

By addressing the whole deep learning pipeline, from the task definition and dataset creation to the model training and deployment, the research conducted underscores potential avenues for addressing the challenge of automatic sensitive content detection, particularly with the use of a customised loss function to address the problem of dataset imbalance, albeit with certain simplifications necessitated by the constraints of the image classification task: while binary classification proved feasible for distinguishing between sensitive and non-sensitive images, it also revealed inherent limitations, particularly in capturing nuanced aspects of colonial oppression. The complexity of this task warrants consideration of alternative approaches such as object detection and leveraging archival metadata for a more comprehensive interpretation. Moreover, other more technical shortcomings emerged from the error analysis conducted on the predictions of the final deployed model: issues like intra-class variation, cluttered environments, and limited high-quality training data pose significant hurdles to accurate classification. Furthermore, the subjective nature of defining sensitivity in colonial archives stresses the need for interdisciplinary collaboration and a nuanced understanding of historical contexts.

In conclusion, this research serves as a foundational step towards a deeper exploration of sensitive content in colonial photographic archives but also underscores the complexity and sensitivity intrinsic in navigating these controversial archives. While the results of the ML experiments demonstrate promising outcomes in detecting sensitive content, there are inherent trade-offs in algorithmic simplification. Addressing the identified limitations requires a multifaceted interdisciplinary approach, encompassing stakeholder engagement, utilisation of metadata, and the exploration of advanced multimodal ML techniques such as GLIP and CLIP which hold promise in capturing the rich interplay between visual and textual archival elements.

Appendix A

On the use of language

The ongoing international debate on the decolonisation of cultural institutions in the GLAM sector has lead to a push for diversity in traditional structures and collections, often resulting in the presentation of usage guidelines for both the general public and professionals.

In the case of archives, the language used is a common topic of debate, as it is often the remnants of surpassed views that have not been updated. The connection between traditional archival practices and the Eurocentric colonial ideologies of the past can result in archival descriptions that use offensive terms [Chilcott, 2019]. Several resources have been produced to promote more awareness on the role of archivists in the endorsement of distressing traditional archival procedures and the employment of a behaviour that is as respectful as possible towards the hurtful records used. These tools were used throughout the whole pipeline of this research along with continuous discussions on the terminology used and include: Archives for Black Lives in Philadelphia's *Anti-racist description resources*, a compendium of metadata recommendations for addressing and countering traditional racist archival descriptions, that helps professionals in the field to understand the issue of describing underrepresented and marginalised groups [Antracoli et al., 2020]; and "Woorden doen ertoe" (Words Matter), a publication compiled by the National Museum of World Cultures (formerly known as Tropenmuseum, Afrika Museum, Museum Volkenkunde, Wereldmuseum), which provides guidance on the use of words and explains possible sensitivities and alternatives. These studies are openly iterative and *work-in-progress*, positioning themselves as an active part of the ongoing discussion on sensitivity in the GLAM sector and encouraging feedback from other professionals and users.

We deem it sensible to explain the use of specific terminology, used especially in Chapter 3:

1. "Indigenous", used instead of "native"
   (a) The term "Indigenous" has been preferred due to the problematic colonial implications of the word "native", which traditionally posits a colonial hierarchy.
   (b) As this research has a broad geographical scope, the term "Indigenous" has been used in its generic connotation to describe the various peoples colonised by European countries.
2. "Race", used instead of "ethnicity"
   (a) The term "race" has been preferred to "ethnicity" in order to recognise and make explicit the implied biological differences between different groups of people: while "ethnicity" refers more to shared cultural practices, the concept of "race" is based on physical characteristics (e.g. skin colour), which have been used to create a hierarchical categorisation of humans, reinforcing colonial ideologies.
   (b) The significant social consequences of this idea have made it important to use the term, albeit in inverted commas to emphasise its problematic nature.

Another fundamental aspect throughout the development of this thesis has been that of the scholars in the field of "ethics of care", especially when involving archival institutions: there is a need to recognise the importance of the ethics of archival research throughout the whole project cycle, and even more so when researching human suffering [Subotić, 2020], and to prioritise the needs of users as emotional beings over the all-encompassing "neutrality" of archives, in order to open up a field that has conventionally been perceived as unwelcoming and inaccessible [Miller, 2021, Caswell, 2014].

Bibliography

[Agarap, 2019] Agarap, A. F. (2019). Deep learning using rectified linear units (relu).

[Agostinho, 2019] Agostinho, D. (2019). Archival encounters: rethinking access and care in digital colonial archives. *Archival Science*, 19(2):141–165.

[Ali et al., 2023] Ali, D., Milleville, K., Verstockt, S., Van de Weghe, N., Chambers, S., and Birkholz, J. M. (2023). Computer vision and machine learning approaches for metadata enrichment to improve searchability of historical newspaper collections. *Journal of Documentation*.

[Anderson, 2013] Anderson, J. (2013). *Anxieties of authorship in the colonial archive*, pages 229–246. Taylor and Francis.

[Antracoli et al., 2020] Antracoli, A. A., Berdini, A., Bolding, K., Charlton, F. C., and Ferrara, A. (2020). Archives for black lives in philadelphia: Anti-racist description resources. *Antiracism Digital Library*, 6.

[Arnold and Tilton, 2022] Arnold, T. and Tilton, L. (2022). *Analyzing Audio/Visual Data in the Digital Humanities*, chapter 17, page 179–187. Bloomsbury Academic.

[Arnold and Tilton, 2023] Arnold, T. and Tilton, L. (2023). *Distant Viewing: Computational Exploration of Digital Images*. The MIT Press.

[Aske and Giardinetti, 2023] Aske, K. and Giardinetti, M. (2023). (mis)matching metadata: Improving accessibility in digital visual archives through the eycon project. *J. Comput. Cult. Herit.*, 16(4).

[Azoulay, 2019] Azoulay, A. A. (2019). *Potential history: Unlearning Imperialism*. Verso.

[Bailey, 1994] Bailey, K. D. (1994). *Typologies and taxonomies: An introduction to classification techniques*, volume 102. Sage.

[Bocyte. and Oomen., 2020] Bocyte., R. and Oomen., J. (2020). Content adaptation, personalisation and fine-grained retrieval: Applying ai to support engagement with and reuse of archival content at scale. In *Proceedings of the 12th International Conference on Agents and Artificial Intelligence - Volume 1: ARTIDIGH*, pages 506–511. INSTICC, SciTePress.

[Bolukbasi et al., 2016] Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

[Borrini, 2024] Borrini, O. M. (2024). 'Revealing contested memory' Master Degree thesis - Github repository.

[Brecke, 1997] Brecke, P. (1997). Beyond typologies: A taxonomy of violent conflicts. In *Annual Meeting of the International Studies Association, Toronto*.

[Brown, 2018] Brown, C. (2018). *Archival futures*. Facet Publishing.

[Buolamwini and Gebru, 2018] Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.

[Caliskan et al., 2017] Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

[Candela et al., 2023] Candela, G., Pereda, J., Sáez, D., Escobar, P., Sánchez, A., Torres, A. V., Palacios, A. A., McDonough, K., and Murrieta-Flores, P. (2023). An ontological approach for unlocking the colonial archive. *J. Comput. Cult. Herit.*, 16(4).

[Caswell, 2014] Caswell, M. (2014). Seeing Yourself in History: Community Archives and the Fight Against Symbolic Annihilation. *The Public Historian*, 36(4):26–37.

[Chilcott, 2019] Chilcott, A. (2019). Towards protocols for describing racially offensive language in uk public archives. *Archival Science*, 19(4):359–376.

[Colavizza et al., 2021] Colavizza, G., Blanke, T., Jeurgens, C., and Noordegraaf, J. (2021). Archives and ai: An overview of current debates and future perspectives. *J. Comput. Cult. Herit.*, 15(1).

[Colavizza et al., 2019] Colavizza, G., Ehrmann, M., and Bortoluzzi, F. (2019). Index-driven digitization and indexation of historical archives. *Frontiers in Digital Humanities*, 6:4.

[Cole, 2019] Cole, T. (2019). When the camera was a weapon of imperialism. (and when it still is.).

[Cook, 2001] Cook, T. (2001). Fashionable nonsense or professional rebirth: postmodernism and the practice of archives. *Archivaria*, pages 14–35.

[Council et al., 2004] Council, N. R. et al. (2004). *Measuring racial discrimination*. National Academies Press.

[Crane, 2008] Crane, S. A. (2008). Choosing not to look: Representation, repatriation, and holocaust atrocity photography. *History and Theory*, 47(3):309–330.

[De Wilde and Hengchen, 2017] De Wilde, M. and Hengchen, S. (2017). Semantic enrichment of a multilingual archive with linked open data. *Digital Humanities Quarterly*.

[Deng et al., 2009] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

[di Lenardo et al., 2016] di Lenardo, I., Seguin, B., and Kaplan, F. (2016). Visual patterns discovery in large databases of paintings. In *Digital Humanities Conference*.

[D'ignazio and Klein, 2023] D'ignazio, C. and Klein, L. F. (2023). *Data feminism*. MIT press.

[Dosovitskiy et al., 2020] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

[Fiorucci et al., 2020] Fiorucci, M., Khoroshiltseva, M., Pontil, M., Traviglia, A., Del Bue, A., and James, S. (2020). Machine learning for cultural heritage: A survey. *Pattern Recognition Letters*, 133:102–108.

[Foliard, 2021] Foliard, D. (2021). Colonialism and its regimes of visibility: Edgard imbert's views of the french empire. *The Journal of Imperial and Commonwealth History*, 49(2):223–259.

[Galar et al., 2012] Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., and Herrera, F. (2012). A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4):463–484.

[Greenhalgh, 2004] Greenhalgh, M. (2004). *Art History*, chapter 3, pages 31–45. John Wiley & Sons, Ltd.

[Gupta and Kapoor, 2020] Gupta, A. and Kapoor, N. (2020). Comprehensiveness of archives: A modern ai-enabled approach to build comprehensive shared cultural heritage. *arXiv preprint arXiv:2008.04541*.

[Han et al., 2021] Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., Qiu, J., Yao, Y., Zhang, A., Zhang, L., Han, W., Huang, M., Jin, Q., Lan, Y., Liu, Y., Liu, Z., Lu, Z., Qiu, X., Song, R., Tang, J., Wen, J.-R., Yuan, J., Zhao, W. X., and Zhu, J. (2021). Pre-trained models: Past, present and future. *AI Open*, 2:225–250.

[He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

[Huang et al., 2016] Huang, C., Li, Y., Loy, C. C., and Tang, X. (2016). Learning deep representation for imbalanced classification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5375–5384.

[ICA, 2012] ICA, I. C. o. A. (2012). Principles of access to archives.

[Jeurgens and Karabinos, 2020] Jeurgens, C. and Karabinos, M. (2020). Paradoxes of curating colonial memory. *Archival Science*, 20(3):199–220.

[Krizhevsky et al., 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

[Kundisch et al., 2021] Kundisch, D., Muntermann, J., Oberländer, A. M., Rau, D., Röglinger, M., Schoormann, T., and Szopinski, D. (2021). An update for taxonomy designers: Methodological guidance from information systems research. *Business & Information Systems Engineering*, 64(4):421–439.

[Li et al., 2021] Li, L. H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J., Chang, K., and Gao, J. (2021). Grounded language-image pre-training. *CoRR*, abs/2112.03857.

[Loebig, 2023] Loebig, L. (2023). The new archivist. Accessed on January 28, 2024.

[Lum and Isaac, 2016] Lum, K. and Isaac, W. (2016). To Predict and Serve? *Significance*, 13(5):14–19.

[Luthra et al., 2022] Luthra, M., Todorov, K., Jeurgens, C., and Colavizza, G. (2022). Unsilencing colonial archives via automated entity recognition.

[Macias, 2016] Macias, T. (2016). Between violence and its representation: Ethics, archival research, and the politics of knowledge production in the telling of torture stories. *Intersectionalities: A Global Journal of Social Work Analysis, Research, Polity, and Practice. Vol 5 No 1. Special Issue: The Ethics and Politics of Knowledge Production. Guest Edited by Teresa Macias*, 5.

[Macknight, 2011] Macknight, E. C. (2011). Archives, heritage, and communities. *Historical Reflections/Reflexions Historiques*, 37(2).

[Manžuch, 2017] Manžuch, Z. (2017). Ethical issues in digitization of cultural heritage. *Journal of Contemporary Archival Studies*, 4(2):4.

[Marciano et al., 2018] Marciano, R., Lemieux, V., Hedges, M., Esteva, M., Underwood, W., Kurtz, M., and Conrad, M. (2018). *Chapter 9: Archival Records and Training in the Age of Big Data*, page 179–199. Emerald Publishing Limited.

[Marden et al., 2013] Marden, J., Li-Madeo, C., Whysel, N., and Edelstein, J. (2013). Linked open data for cultural heritage: evolution of an information technology. In *Proceedings of the 31st ACM International Conference on Design of Communication*, SIGDOC '13, page 107–112, New York, NY, USA. Association for Computing Machinery.

[Mignolo and Walsh, 2018] Mignolo, W. D. and Walsh, C. E. (2018). *On decoloniality: Concepts, analytics, praxis*. Duke University Press.

[Miller, 2021] Miller, P. (2021). An ethic of care: Interrogating the need for care in the archives. *The iJournal: Graduate Student Journal of the Faculty of Information*, 6(2).

[Mitchell, 1997] Mitchell, T. M. (1997). Machine learning.

[Mitchell, 2005] Mitchell, W. J. (2005). There are no visual media. *Journal of visual culture*, 4(2):257–266.

[Mohammed et al., 2016] Mohammed, M., Khan, M. B., and Bashier, E. B. M. (2016). *Machine Learning*. CRC Press.

[Moretti, 2013] Moretti, F. (2013). Distant reading. *Londyn: Verso Books*.

[Moss et al., 2018] Moss, M., Thomas, D., and Gollins, T. (2018). The reconfiguration of the archive as data to be mined. *Archivaria*, 86(86):118–151.

[Mulya and Bramantya, 2023] Mulya, L. and Bramantya, A. R. (2023). Problems in 'accessing' colonial archives for indonesian history department student. In *Proceedings of the International Joint Conference on Arts and Humanities 2022 (IJCAH 2022)*, pages 1595–1600. Atlantis Press.

[Noble, 2018] Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.

[Noord et al., 2021] Noord, N., Olesen, C., Ordelman, R., and Noordegraaf, J. (2021). Automatic annotations and enrichments for audiovisual archives. In Rocha, A., Steels, L., and van den Herik, J., editors, *Proceedings of the 13th International Conference on Agents and Artificial Intelligence - Volume 1: ARTIDIGH*, pages 633–640. SciTePress.

[Odumosu, 2020] Odumosu, T. (2020). The crying child: On colonial archives, digitization, and ethics of care in the cultural commons. *Current Anthropology*, 61(S22):S289–S302.

[Omi and Winant, 2014] Omi, M. and Winant, H. (2014). *Racial formation in the United States*. Routledge, London, England, 3 edition.

[Osterhammel, 2005] Osterhammel, J. (2005). *Colonialism: A Theoretical Overview*. Markus Wiener Publishers.

[Prabhu and Birhane, 2020] Prabhu, V. U. and Birhane, A. (2020). Large image datasets: A pyrrhic win for computer vision?

[Radford et al., 2021] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020.

[Ranade, 2016] Ranade, S. (2016). Traces through time: A probabilistic approach to connected archival data. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 3260–3265. IEEE.

[Reed, 2021] Reed, T. (2021). Indigenous dignity and the right to be forgotten. *Brigham Young University Law Review*.

[Risam, 2018] Risam, R. (2018). *New digital worlds: Postcolonial digital humanities in theory, praxis, and pedagogy*. Northwestern University Press.

[Rolan et al., 2019] Rolan, G., Humphries, G., Jeffrey, L., Samaras, E., Antsoupova, T., and Stuart, K. (2019). More human than human? artificial intelligence in the archive. *Archives and Manuscripts*, 47(2):179–203.

[Romein et al., 2020] Romein, C. A., Kemman, M., Birkholz, J. M., Baker, J., De Gruijter, M., Meroño-Peñuela, A., Ries, T., Ros, R., and Scagliola, S. (2020). State of the field: Digital history. *History*, 105(365):291–312.

[Said, 1995] Said, E. W. (1995). *Orientalism*. Penguin.

[Sarker, 2021] Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, 2(3).

[Schwartz et al., 2022] Schwartz, R., Vassilev, A., Greene, K. K., Perine, L., Burt, A., and Hall, P. (2022). Towards a standard for identifying and managing bias in artificial intelligence.

[Smeulders et al., 2000] Smeulders, A. W., Worring, M., Santini, S., Gupta, A., and Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transactions on pattern analysis and machine intelligence*, 22(12):1349–1380.

[Smits and Wevers, 2022] Smits, T. and Wevers, M. (2022). The agency of computer vision models as optical instruments. *Visual Communication*, 21(2):329–349.

[Snydman et al., 2015] Snydman, S., Sanderson, R., and Cramer, T. (2015). The international image interoperability framework (iiif): A community & technology approach for web-based images. *Archiving Conference*, 12(1):16–21.

[Sontag, 1979] Sontag, S. (1979). *On Photography*. Penguin.

[Stoler, 2002] Stoler, A. L. (2002). Colonial archives and the arts of governance. *Archival Science*, 2(1–2):87–109.

[Subotić, 2020] Subotić, J. (2020). Ethics of archival research on political violence. *Journal of Peace Research*, 58(3):342–354.

[Tharwat, 2020] Tharwat, A. (2020). Classification assessment methods. *Applied Computing and Informatics*, 17(1):168–192.

[Trouillot, 2015] Trouillot, M.-R. (2015). *Silencing the past: Power and the production of history*. Beacon Press.

[UNESCO, 2003] UNESCO (2003). Recommendation concerning the promotion and use of multilingualism and universal access to cyberspace.

[Upward et al., 2019] Upward, F., Reed, B., Oliver, G., and Clayton, J. E. (2019). 16 recordkeeping informatics for a networked age. *The American Archivist*, 82(1).

[van Noord, 2022] van Noord, N. (2022). A survey of computational methods for iconic image analysis. *Digital Scholarship in the Humanities*, 37(4):1316–1338.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

[Vellido, 2019] Vellido, A. (2019). The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Computing and Applications*, 32(24):18069–18083.

[Verstockt et al., 2018] Verstockt, S., Nop, S., Vandecasteele, F., Baert, T., Van de Weghe, N., Paulussen, H., Rizza, E., and Roeges, M. (2018). Ugesco - a hybrid platform for geo-temporal enrichment of digital photo collections based on computational and crowdsourced metadata generation. In Ioannides, M., Fink, E., Brumana, R., Patias, P., Doulamis, A., Martins, J., and Wallace, M., editors, *Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection*, pages 113–124, Cham. Springer International Publishing.

[Wang et al., 2017] Wang, Y.-X., Ramanan, D., and Hebert, M. (2017). Learning to model the tail. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

[Wevers and Smits, 2019] Wevers, M. and Smits, T. (2019). The visual digital turn: Using neural networks to study historical images. *Digital Scholarship in the Humanities*, 35(1):194–207.

[Yakel, 2003] Yakel, E. (2003). Archival representation. *Archival Science*, 3:1–25.

[Zhu et al., 2023] Zhu, H., Chen, B., and Yang, C. (2023). Understanding why vit trains badly on small datasets: An intuitive perspective. *arXiv preprint arXiv:2302.03751*.

[Zhuang et al., 2021] Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q. (2021). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76.