

Don't Erase, Inform! Detecting and Contextualizing Harmful Language in Cultural Heritage Collections

Orfeas Menis Mastromichalakis¹ Jason Liartis¹ Kristina Rose²
Antoine Isaac^{3,4} and Giorgos Stamou¹

¹ National Technical University of Athens ² DFF - Deutsches Filminstitut & Filmmuseum

³ Europeana Foundation ⁴ Vrije Universiteit Amsterdam
{menorf, jliartis}@ails.ece.ntua.gr, rose@dff.film
antoine.isaac@europeana.eu, gstam@cs.ntua.gr

Abstract

Cultural Heritage (CH) data hold invaluable knowledge, reflecting the history, traditions, and identities of societies, and shaping our understanding of the past and present. However, many CH collections contain outdated or offensive descriptions that reflect historical biases. CH Institutions (CHIs) face significant challenges in curating these data due to the vast scale and complexity of the task. To address this, we develop an AI-powered tool that detects offensive terms in CH metadata and provides contextual insights into their historical background and contemporary perception. We leverage a multilingual vocabulary co-created with marginalized communities, researchers, and CH professionals, along with traditional NLP techniques and Large Language Models (LLMs). Available as a standalone web app and integrated with major CH platforms, the tool has processed over 7.9 million records, contextualizing the contentious terms detected in their metadata. Rather than erasing these terms, our approach seeks to inform, making biases visible and providing actionable insights for creating more inclusive and accessible CH collections.

1 Introduction

Language is central to Cultural Heritage, shaping how professionals and the public interpret historical collections. However, the descriptions of CH objects can contain outdated or offensive language that reflect historical cultural and societal norms and power structures no longer appropriate today. While CHIs recognize this issue, updating legacy metadata is a daunting task due to the sheer scale of their collections and the complexity of the process.

Artificial Intelligence (AI) offers scalable solutions by performing sophisticated tasks and processing large volumes of data. However, AI systems can inadvertently perpetuate biases inherent in the training data or developers' assumptions.

To mitigate this, human-in-the-loop approaches are crucial, integrating input from domain experts, marginalized communities, and CH professionals to ensure both accuracy and cultural sensitivity.

With this work, we aim to improve CH metadata, by making existing bias visible, guiding users in understanding the original context of words, and how they are currently perceived. To do so, we introduce a multilingual vocabulary and an AI-powered tool for detecting offensive terms in CH metadata, providing contextual explanations and suggested alternatives. The vocabulary, co-created with marginalized communities, researchers, and CH professionals, pairs biased language with contextual information and suggestions for the appropriate usage of terms. It is structured as a Knowledge Graph (KG) following the approach of [Nesterov et al. \(2023\)](#) and standards like the Simple Knowledge Organization System (SKOS) ([Baker et al., 2013](#)). Both the created vocabulary¹, and the respective ontology² describing its structure, are available on EU Vocabularies³, a platform by the Publications Office of the EU that provides controlled vocabularies and reference data. This publication enhances the vocabulary's reach and facilitates its integration into broader CH initiatives. The system was evaluated through a series of human-in-the-loop validation campaigns refining both the tool and vocabulary. The resulting dataset from these campaigns is published as a benchmark for derogatory term detection in CH.

To ensure accessibility, we developed a web application⁴ that allows manual text input, batch processing, and exploration of results. Additionally,

¹<https://op.europa.eu/en/web/eu-vocabularies/dataset/-/resource?uri=http://publications.europa.eu/resource/dataset/de-bias-vocabulary>

²<https://op.europa.eu/de/web/eu-vocabularies/dataset/-/resource?uri=http://publications.europa.eu/resource/dataset/de-bias-ontology>

³<https://op.europa.eu/en/web/eu-vocabularies>

⁴<https://debias-tool.ails.ece.ntua.gr/>

we provide an API for integration. Through this API, the tool has been deployed on Europeana⁵, the European digital heritage platform, and CH metadata ingestion platforms like MINT⁶, processing over 7.9 million records. A dedicated UI component in Europeana now informs users about detected contentious terms. The resources and code used throughout this paper can be found in our GitHub repository⁷.

2 Related Work

Our work relates to research on toxic, abusive, and offensive language, as well as hate speech. Prior efforts have sought to establish a unified taxonomy of harmful content to enable structured analysis (Banko et al., 2020; Kurrek et al., 2020). In this work, we use the umbrella terms hateful or harmful language (European Commission et al., 2023).

Detecting harmful language is a critical challenge that was traditionally handled manually by moderators, administrators, or curators. However, the vast volume of content now demands automated solutions (Waseem and Hovy, 2016). Research in this area spans various domains, including workplace communications (Bhat et al., 2021), online chatbots (Song et al., 2024), social media (Niraula et al., 2021), and news articles (Subbiah et al., 2023). Early methods relied on SVMs (Warner and Hirschberg, 2012), often enhanced with curated dictionaries (Tulkens et al., 2016) and ensemble models (Burnap and Williams, 2014). Later, deep learning techniques, such as CNNs and LSTMs, became popular (Del Vigna et al., 2017; Mathur et al., 2018; Meyer and Gambäck, 2019; Chakrabarty et al., 2019; Modha et al., 2018). Recent advances have focused on transformer-based (Elmadany et al., 2020; Alonso et al., 2020; Davidson et al., 2020) and LLM-based approaches (Yao et al., 2024; Plaza-del Arco et al., 2023).

However, harmful language datasets have been shown to suffer from robustness issues (Korre et al., 2023) and inherent biases (Wiegand et al., 2019; Davidson et al., 2019; Xia et al., 2020; Zhang et al., 2020). Similarly, data-driven methods often exhibit limitations in robustness (Kaushik et al., 2020; Sen et al., 2022) and racial biases (Sap et al., 2019), raising concerns about their fairness, real-world applicability, and potential to perpetuate harmful biases

(Thiago et al., 2021). Context is essential in detecting harmful language (Bourgeade et al., 2024). However, many detection models struggle to account for context, leading to high false-positive rates, as noted in (Kennedy et al., 2020). Our work addresses this issue through an LLM-based disambiguation module that leverages a curated vocabulary, explicitly defining the context in which a term is considered derogatory. Another key aspect of our approach is the involvement of affected communities in term collection, context definition, and system evaluation. This aligns with recent calls for equitable representation in such tasks (Kurrek et al., 2020) and research emphasizing the importance of community context (Saleem et al., 2022) and cultural nuances in harmful language detection (Zhou et al., 2023).

The CH domain presents unique challenges that demand a more nuanced approach. Unlike user-generated content, CH metadata references historical entities, events, and societal norms that reflect the language of their time. Simply removing or censoring problematic terms poses risks akin to erasing parts of history and limiting opportunities for critical engagement. Instead, harmful language detection in CH must balance historical accuracy with ethical concerns, ensuring metadata remains both informative and sensitive to evolving societal values.

In CH, efforts to address contentious terminology have built upon glossaries of problematic terms in museum databases (Modest, 2018) and evolved into knowledge graphs that encode expert knowledge for more inclusive digital collections (Nesterov et al., 2023). We build on this by extending their proposed KG schema, enriching the vocabulary with additional terms, and expanding coverage beyond English and Dutch.

3 The Vocabulary

The vocabulary is a comprehensive resource designed to identify contentious terms, explain their historical and contextual significance, and guide appropriate usage. It powers the harmful language detection tool (see Section 4), enabling the identification of outdated or potentially harmful language in CH metadata. Additionally, it serves as a standalone reference for CH professionals, and curators to review and refine collection descriptions.

⁵<https://www.europeana.eu/>

⁶<https://mint-projects.image.ntua.gr/mint4all>

⁷<https://github.com/ails-lab/de-bias>

3.1 Co-creating the Vocabulary

The vocabulary includes contentious terms in English, French, German, Dutch, and Italian that are known or likely to appear in CH metadata. It primarily focuses on ‘Ethno-religious identity’, ‘Gender and sexual identity’, and ‘Migration and colonial past’, with extensions to other domains, such as disability. Each term is accompanied by contextual explanations of its problematic nature and recommendations for appropriate handling. The vocabulary was developed iteratively, grounded in a typology of bias patterns (Masschelein et al., 2024) that categorizes bias by factors such as religion, sexuality, and ethnicity through an intersectional lens (Crenshaw, 2013). This framework was combined with an analysis of linguistic patterns, including euphemisms, dysphemisms, and diminutive, that perpetuate bias in CH metadata. Editorial guidelines ensured consistency and usability, recommending lower-secondary school reading levels for contextual descriptions, avoiding Eurocentric generalizations, and providing clear recommendations. Additionally, to support UI integration and effectively inform users about contentious terms, we included recommendations on the format and length of descriptions for contentious terms and issues, ensuring compatibility with interfaces that adhere to established UI standards.

The vocabulary compilation followed a twofold approach: first, we consulted existing glossaries, research publications, encyclopedias, and dictionaries. Scientific literature, along with glossaries, informed contextual descriptions covering etymology, historical usage, and debates around the selected terms. We also examined term occurrences in CH platforms like Europeana to assess their relevance. These sources guided recommendations for handling contentious terms or suggesting alternatives. A full list of the 170+ sources used is available in our GitHub repository⁸. Second, terms and descriptions were collected, reviewed, and refined through 12 co-creation sessions, which served as safe spaces for debating and discussing bias issues with affected communities and their allies. The sessions were language-specific and focused on distinct topics: (i) Ethnicity and ethno-religious identity with Jewish communities (German), (ii) Gender and sexual identity with LGBTQIA+ communities (Italian, English), and (iii) Migration and

⁸https://github.com/ails-lab/de-bias/blob/main/resources/vocabulary_sources.csv

colonial history with Dutch- and French-speaking participants. Only terms relevant to the community’s thematic focus were discussed in depth during each session. In total, over 60 community members and allies contributed. We refer to the *Ethical Considerations* section at the end of this paper for more details about the support and remuneration of the participants of these events who worked on this sensitive topic.

To guide and evaluate community engagement, we developed a methodology⁹ focused on building and maintaining relationships with communities while ensuring a respectful and inclusive approach. Although specific setups and tools were not predetermined, three primary interaction patterns emerged: (i) Participants searched CH platforms (e.g. Europeana.eu, Deutsche Digitale Bibliothek¹⁰) and institutional databases for contentious terms—either pre-selected by the editorial team or suggested by the community—then debated their findings. (ii) Existing CH object descriptions were reviewed for accuracy, inclusiveness, and positionality. Discussions explored whether descriptions should retain historical language or take a more interpretive approach to balance objectivity and subjectivity. (iii) In follow-up meetings, terms and contextual descriptions were revisited for further discussion, allowing for iterative refinement. These sessions offered valuable insights into community perspectives on contextualization. Participants emphasized that beyond identifying harmful language, CH metadata often suffers from poor representativeness, failing to reflect the viewpoints and knowledge of the communities. Many highlighted the importance of contextualizing contentious terms rather than merely replacing or masking them, as this approach preserves historical accuracy, acknowledges past discriminations, and recognizes the resilience of those who resisted them.

3.2 Challenges

Several challenges emerged during the development of the vocabulary. One major challenge was its multilingual nature, as bias does not directly translate across languages. For example, in English, ‘queer’ once had a negative connotation before being reclaimed by the LGBTQIA+ commu-

⁹https://pro.europeana.eu/files/Europeana_Professional/Projects/debias/a_community_engagement_methodology_resources_reflections_recommendations_v3_july_2024.pdf

¹⁰www.deutsche-digitale-bibliothek.de

nity, whereas in Italian, it was introduced more recently with a solely positive meaning. Consequently, ‘queer’ was included in the English subset but not the Italian one. Similarly, terms with equivalents across languages required language-specific contextualization. To address these discrepancies, we developed independent language-specific subsets, enabling editors and community participants to focus on their respective linguistic contexts while optionally cross-referencing other languages. This approach reduced editorial complexity and allowed more thorough research and community engagement. While terms were not forcibly unified across languages at this stage, their semantic relationships are represented in the knowledge graph (see Section 3.3). For example, ‘Third World’ (English), ‘Derde Wereld’ (Dutch), and ‘Dritte Welt’ (German) are distinct entries linked semantically. Although some links have been established, further refinement will be pursued in future iterations, ensuring an iterative and flexible process without compromising expressivity. This approach also accommodated imbalances in thematic coverage across languages, stemming from the focus areas of community engagement. For instance, the Dutch vocabulary emphasizes migration and the colonial past, reflecting the priorities of Dutch-speaking participants, while the German subset focuses more on Ethno-religious identity due to its emphasis on antisemitism. Additionally, differences in existing research on diversity, inclusion, and discrimination across countries and languages further contributed to variations in vocabulary size and focus.

Another set of challenges pertained to describing the contentiousness of terms. First, historical accuracy was crucial, as CH metadata often references historical artifacts or language from past decades or centuries. However, meanings evolve over time, requiring descriptions that reflect these shifts while enabling consistent flagging, regardless of data’s creation date, which is often unknown. Second, some terms are used in both contentious and neutral contexts, necessitating careful handling to avoid false positives. Approximately 25% of the collected terms exhibit such ambiguity. For example, ‘race’ in English can refer to the contested concept of human races or to a competitive event. To address this, we implemented a filtering mechanism in the harmful language detection tool that assesses contextual usage before flagging a term. During vocabulary development, ambigu-

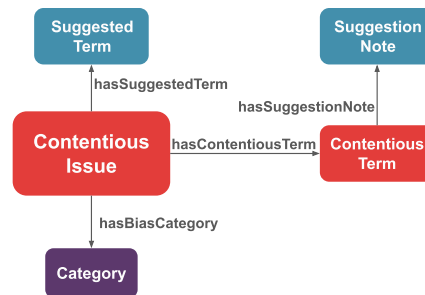


Figure 1: A simplified version of the KG structure

ous terms were explicitly marked, and descriptions were structured to highlight problematic uses while acknowledging neutral ones where necessary. By tackling these challenges, the vocabulary maintains linguistic sensitivity, historical accuracy, and contextual precision, ensuring effective detection of harmful terms across diverse CH metadata.

3.3 Structure and Statistics

To ensure interoperability and computational accessibility, we structured our vocabulary as a Knowledge Graph (KG) based on the SKOS standard (Baker et al., 2013). This design enables a structured representation of terms and their relationships while ensuring scalability and flexibility for future updates. Our approach builds on prior work (Nesterov et al., 2023), which we extended and adapted to our specific needs. At the core of the KG introduced in Nesterov et al. (2023) are contentious terms, their associated contentious issues, suggested alternatives, and suggestion notes. The ContentiousTerm entity represents each term in the vocabulary, while ContentiousIssue provides contextual information, including etymology, relevant discourse, and references explaining why a term (or set of terms) is considered problematic. SuggestionNotes offer standardized and multilingual recommendations on appropriate usage, and SuggestedTerms provide alternative wording. Synonyms and different terms within the same context are grouped by linking multiple ContentiousTerms to the same ContentiousIssue, while semantic matching across languages is achieved with the appropriate SKOS properties.

We slightly modified the linking between contentious issues, suggestions and labels and extended their structure in several ways. Firstly, we classified the issues using bias categories, as described in Section 3.1, directly linking them to the relevant contentious issues. To support harmful lan-

guage detection tools, ambiguous terms are flagged for contextual analysis. We also grouped gendered forms, and spelling variations by linking them to the same ContentiousIssue similarly to how synonyms were treated. Finally, our KG also includes sources and administrative metadata, such as versioning, modification history, and rights information, as a means to foster sustainable reuse of our data by consuming services. Figure 1 illustrates a simplified version of our KG structure adapted from Nesterov et al. (2023). For further details on the knowledge graph structure, we refer the reader to DE-BIAS Project (2024).

Currently, the vocabulary consists of 687 contentious terms linked to 530 distinct contentious issues. The English subset contains 220 terms, followed by German (163), Dutch (161), French (75), and Italian (68). The vocabulary has been made publicly available through its publication to the EU-Vocabularies, ensuring ongoing maintenance and enhancing its reach and broader adoption. In Appendix A we provide an example of a vocabulary term.

4 The tool

Figure 2 illustrates the tool’s architecture, designed to detect contentious terms leveraging the structured vocabulary described in Section 3. First, inputs undergo tokenization, followed by compound word splitting for German and Dutch, enabling the detection of contentious terms within arbitrarily formed compounds. Next, a lemmatization module standardizes word forms, ensuring accurate matching. The lemmatized texts are then processed independently by two modules: Named Entity Recognition (NER), which identifies proper names, and string matching, which detects vocabulary terms within the input text. These outputs are then merged in a filtering module, which discards detections that are part of named entities. For terms flagged as ambiguous in the vocabulary, an LLM-based disambiguation module is applied. This module evaluates the surrounding context, alongside the contentious issue description and suggestion notes, to determine whether a term is used in a derogatory manner and should be flagged. The final output consists of detected contentious terms within the input text. The following subsections provide a detailed breakdown of the main components.

4.1 Compound Word Splitting

In German and Dutch, it is common to arbitrarily form compound words, combining multiple terms into a single unit. To detect contentious terms in such cases, we developed a compound word splitter, applied before lemmatization to ensure that the lemmatizer processes individual word components. For German, we adapted an open-source compound splitter¹¹, which relies on an external dictionary¹² and grammatical rules for pluralization to correctly decompose complex compounds. Since fewer resources exist for Dutch, we modified the German tool by replacing the dictionary with one from OpenOffice¹³, adjusting pluralization rules to fit Dutch morphology, and refining the splitting logic to avoid over-splitting, which initially led to incorrect matches.

4.2 Lemmatization and String Matching

To detect contentious terms in text we combine lemmatization with exact string matching. Lemmatization maps words to their canonical (dictionary) form, enabling the detection of inflected terms. We employed the lemmatizers from Stanford Stanza (Qi et al., 2020), which combine rule-based and machine-learning methods balancing precision and generalization. Since many terms in our vocabulary are antiquated or slurs, they are underrepresented in standard training corpora. To mitigate this, Stanza’s character-level processing enables morphological analysis of words, i.e. analysing their base, inflections to express number or tense, etc., improving generalization to rare and unseen terms. We also leveraged Stanza’s custom rule support, implementing 162 rules to enhance accuracy for vocabulary terms that the default lemmatizer misprocessed. To enhance the tool’s efficiency, all vocabulary terms were pre-processed and stored in lemmatized form.

After lemmatization, we opted for exact string matching over substring or similarity-based methods, because vocabulary terms often appear as substrings in unrelated words. Unlike domains where slurs are frequently obfuscated (e.g., numbers replacing letters), CH texts are well-formed, officially transcribed records without intentional evasion. Therefore, techniques for detecting such variations are unnecessary in our case. In the case of terms consisting of multiple tokens, the term is

¹¹https://github.com/repodiac/german_compound_splitter/

¹²<https://sourceforge.net/projects/germandict/>

¹³<https://wiki.openoffice.org/wiki/Dictionaries>

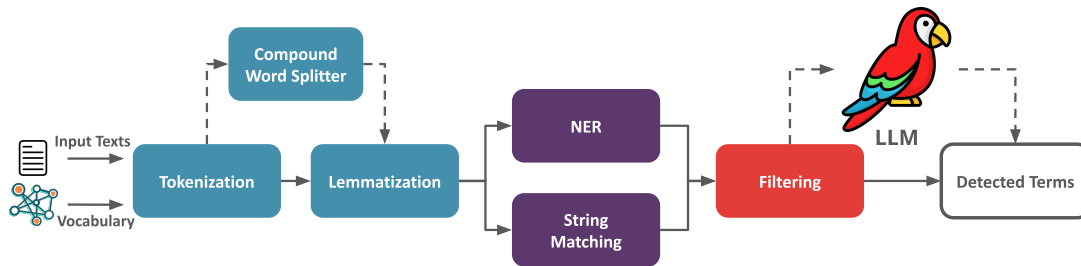


Figure 2: Tool Architecture Overview

only matched with text spans that contain the exact same tokens (after lemmatization) in the exact same order.

4.3 NER

To prevent false detections of vocabulary terms within named entities, we integrated a Named Entity Recognition (NER) module into our pipeline. One common case is terms appearing as surnames, e.g. “Sordo” in Italian. Using the NER module from Stanza, known for its robustness and efficiency, we ensure that terms within named entities are excluded from detection. While the LLM module could also handle this task, relying on it would impact performance due to additional processing time. The dedicated NER component allows us to achieve comparable results efficiently, maintaining both accuracy and speed.

4.4 LLM Disambiguation

Since certain vocabulary terms can be contentious in some contexts while neutral in others, simply flagging all detected instances would overwhelm the annotations with false positives. To address this, we implemented an LLM-based disambiguation module to assess whether a detected term is used in a contentious sense based on its surrounding context. This task is closely related to Word Sense Disambiguation (WSD). Traditional approaches rely on BERT-like encoder models to generate contextual embeddings, which are then classified into predefined senses using either fine-tuned Transformer layers or k-Nearest Neighbors classifiers (Bevilacqua et al., 2021; Loureiro et al., 2021). However, such methods require a large amount of annotated training data, which was not available for most vocabulary terms.

A more suitable approach is the use of LLMs, which have demonstrated strong knowledge acquisition and contextual understanding (Sumanathilaka et al., 2024). Recent research has successfully

leveraged zero-shot and few-shot prompting with LLMs for classification tasks, including WSD (Yae et al., 2024), without requiring fine-tuning. This made them an attractive solution given our need for high adaptability with minimal training data. We tested models of varying sizes and architectures (see Sec 5.2), and conducted multiple rounds of prompt engineering to refine the structure and accuracy of the responses. See Appendix B for further information on the LLMs and the prompts.

5 Evaluation and Discussion

Our evaluation focused primarily on precision to minimize false positives (i.e., incorrectly flagged terms), which can undermine user trust (Sasse, 2015) and acceptance, leading to rejection by users and institutions. A major challenge in evaluating our tool was the lack of ground truth data, as CH metadata annotated for derogatory terms is scarce, if not non-existent. To overcome this, we manually created an evaluation dataset through a series of niche-sourcing campaigns, ensuring a curated and contextually rich set of records. This dataset is also made publicly available as a benchmarking resource for detecting harmful language in CH metadata.

Using the evaluation dataset and a large set of unannotated CH metadata to assess the tool’s computational performance, we evaluated two main aspects: precision and throughput. We also examined the impact of the NER and LLM-based modules on both metrics. Additionally, we assessed precision separately for each language, testing multiple LLMs, to understand performance variations caused by differences in the vocabulary and the technical components, which sometimes do not perform consistently across all languages.

5.1 Creation of the Evaluation Dataset

To construct a dataset suitable for evaluating precision, we needed a representative set of records

containing contentious terms. However, such terms are sparse in CH metadata, making random selection ineffective for achieving sufficient vocabulary coverage. To overcome this, we first searched for records containing vocabulary terms in a pool of 5 million Europeana records, utilizing the lemmatization and string-matching components of our pipeline. This allowed us to extract a subset of texts covering a wide range of vocabulary terms. For terms not found in Europeana, we manually gathered additional texts from other relevant open sources, such as Wikipedia¹⁴, online dictionaries, and other cultural heritage digital platforms besides Europeana. In a few cases where multiple examples were unavailable for a term, we used ChatGPT¹⁵ to generate supplementary examples, which were subsequently reviewed by experts. Instead of replicating Europeana’s term distribution—where some terms are underrepresented—we opted for a more balanced dataset, ensuring all terms were included without over-representing the most frequent ones. This targeted approach enabled effective precision assessment but, due to its non-random nature, did not provide meaningful insights into accuracy or recall.

To validate the detections and create a gold-standard evaluation dataset, we conducted eight niche-sourcing campaigns on CrowdHeritage (Kaldeli et al., 2021; Ralli et al., 2020), a crowdsourcing platform for CH metadata enrichment. Unlike traditional crowdsourcing, niche sourcing involves domain experts, CH professionals, and affected communities as annotators. Participants could upvote or downvote detections, provide feedback, and highlight additional biased or derogatory terms that were not flagged or not included in the vocabulary. In total, 145 participants contributed, evaluating more than 3,000 detections and casting more than 10,000 validation votes. When consensus was unclear (i.e., the absolute difference between upvotes and downvotes was less than two), campaign organizers manually reviewed these cases, considering user feedback on why certain detections were contested.

This process resulted in a curated evaluation dataset of around 3,700 texts covering all vocabulary terms. We publicly release this dataset alongside this work as a benchmarking resource for detecting contentious terms in CH metadata, comple-

¹⁴<https://www.wikipedia.org/>

¹⁵<https://chatgpt.com/>

LLM	NER	Precision↑	Throughput↑
✗	✗	0.70	15112
✗	✓	0.71	15092
✓	✗	0.86	787
✓	✓	0.87	813

Table 1: The effect of NER and LLM on performance.

menting our vocabulary.

5.2 Results and Discussion

To evaluate the tool’s inference speed, we used a collection of unlabelled, de-duplicated texts from 5 million CH records (the collection referenced in 5.1 for the creation of the evaluation dataset), totaling over 690 million characters. Throughput was measured as the average characters per second processed by the tool in five runs conducted in an isolated environment. For precision, we relied on the constructed annotated evaluation dataset. All experiments were performed on a server with a single NVIDIA GeForce RTX 4090 using llama.cpp¹⁶.

We first conducted an ablation study to evaluate the tool’s performance with and without the NER and LLM-based modules, examining their impact on precision and throughput. For this benchmark, we used the Mixtral-8x7B model quantized to 3 bits. As shown in Table 1, the base tool’s text processing speed is very high (15,112 characters per second), with the primary bottleneck being LLM inference. This is partially alleviated by the NER module, which efficiently filters out irrelevant texts without harming precision. Regarding precision, the NER module slightly improves accuracy when added to the base tool, though this increase depends on the presence of named entities in the evaluation dataset—a relatively rare occurrence in our case. However, in CH platforms, there are fields such as creator and contributor, which typically contain named entities. Therefore, we anticipate a more substantial impact from the NER module in real-world applications. As expected, the LLM module significantly enhances precision by filtering out many false positives caused by ambiguous terms, which would otherwise overwhelm the results with incorrect detections.

We also evaluated a variety of LLMs across all languages, with the results presented in Table 2. Due to memory constraints, some LLMs were quan-

¹⁶<https://github.com/ggml-org/llama.cpp>

Model	Precision ↑					Aggregated
	nl	en	fr	de	it	
EuroLLM-9B	0.42	0.74	0.97	0.83	0.97	0.73
Llama-3.1-8B-Q8	0.76	0.83	0.97	0.88	0.97	0.88
Ministral-8B	0.73	0.89	0.97	0.87	0.97	0.88
Mistral-Small-24B-Q8	0.66	0.89	0.97	0.87	0.98	0.86
Mixtral-8x7B-Q3	0.72	0.83	0.97	0.89	0.98	0.87
StableLM-2-1.6B	0.60	0.72	0.96	0.82	0.95	0.80
StableLM-2-12B	0.72	0.82	0.96	0.87	0.97	0.86

Table 2: Precision of the tool, utilizing different LLMs. Bold indicates the highest precision per language, and aggregated.

tized, indicated in the table by the letter ‘Q’ followed by the number of bits. Precision remained consistently high for Italian and French, which we attribute to the low percentage of ambiguous terms for these languages in the vocabulary. In contrast, precision was lower for Dutch, primarily due to the high number of ambiguous Dutch terms combined with the lower performance of LLMs on Dutch—likely a result of limited Dutch data used in LLM pretraining. Overall, Llama-3.1-8B and Ministral-8B achieved the best precision, with Mixtral-8x7B and StableLM-2-12B performing comparably well. In terms of throughput, StableLM-2-1.6B was the fastest model, with Llama-3.1-8B coming second, providing a good trade-off between speed and accuracy.

6 Application & Impact

To maximize the usability and impact of our tool, we developed a public API endpoint that allows integration into other platforms and services. This API is the basis for a stand-alone web application and it has been deployed in two large-scale CH platforms, Europeana and MINT.

6.1 Standalone Web Application

To ensure accessibility for users with little to no technical expertise—an important request from many practitioners in the CH domain—we developed a standalone web application¹⁷ with an intuitive and interactive interface. The web app provides two main functionalities: (1) users can input custom texts and directly view the detected terms alongside contextual explanations from the vocabulary, detailing why each term is considered derogatory; (2) users can analyze larger datasets by

¹⁷<https://debias-tool.ails.ece.ntua.gr/>

uploading a ZIP file containing multiple text files, with the option to receive via email a JSON file with machine-readable annotations and a statistical report of the results. Users can customize the service by (de-)activating the components for NER and LLM-based disambiguation. This enables them to experiment and optimize performance based on their specific needs—prioritizing processing speed or precision. Independently from such practical usage, the application also serves as a key resource for showcasing the tool and engaging with a broader audience about the issues of bias in CH metadata. In earlier workshops, this enabled us to gather feedback that was useful to refine earlier versions of the tool and the vocabulary based on real-world cases.

6.2 Integration with External Platforms

Europeana, a platform providing access to digital heritage across Europe with contributions from over 3,200 CHIs and more than 10,000 daily visits on europeana.eu, has integrated our tool into its metadata ingestion pipeline. This integration allows data providers to analyze their records for contentious terms before publication. Additionally, Europeana has developed a UI component that highlights these terms on item pages, offering contextual explanations sourced from our vocabulary and linking to the terms published in the EU Vocabularies platform for further reference. This functionality can display detections for newly ingested metadata while also surfacing detections from our tool’s application on over 7.9 million legacy metadata records, resulting in more than 77,000 annotations. This large-scale deployment enhances transparency by identifying and contextualizing contentious terms across the platform.

MINT, the Metadata Interoperability Tool, is a service that enables organizations with diverse

data formats to transform and map their metadata into a standardized schema, which can then be exported to an online repository for harvesting by consuming platforms. Currently, MINT facilitates the publication of over 5.2 million CH records from approximately 200 organizations to multiple platforms, including Europeana. Our tool has been integrated into MINT, allowing users to detect contentious terms in their datasets and annotate them. Organizations can analyze their data in an easy and intuitive way via the MINT UI, without requiring technical expertise to interact with an API. Importantly, this integration gives CHIs direct access to the tool’s functionalities within MINT, while also retaining full control over how to handle the results, ensuring they can assess and integrate the tool’s insights according to their specific needs and policies.

7 Conclusions

In this work, we presented a comprehensive approach to detecting and contextualizing harmful language in CH metadata. We developed a rich multilingual vocabulary of contentious terms, integrating it with an AI-based tool designed to identify potentially harmful language while providing contextual explanations. Additionally, we curated a niche-sourced evaluation dataset, enabling a grounded assessment of our tool’s precision across multiple languages. This dataset is publicly available, supporting further research on harmful language detection in CH contexts. Our tool was deployed at scale via a standalone web app, and integrated with major CH platforms.

Future work will focus on expanding the vocabulary to include more terms and extending coverage to additional languages, enhancing its applicability across diverse cultural contexts. We also plan to investigate advanced technical solutions to improve the precision of detection, including the integration of more sophisticated NLP techniques and adaptive models that better capture context and nuanced meanings. These advancements will also aim to increase accuracy while maintaining transparency, ultimately contributing to more inclusive and representative digital heritage platforms.

Limitations

One limitation of our approach emerged during our community discussions. Participants pointed out a persistent bias in the form of information

omissions, leading to incomplete representations of their culture and history. This issue was particularly evident in collections related to gender and sexual identity (e.g., excluding information about LGBTQIA+ relationships) and migration and colonial histories (e.g., using inaccurate or outdated vocabulary to describe cultural activities or circumstances). Often, these CH objects were described from an uninformed and biased perspective, outside the community. Similarly, a general lack of adequate and representative language in the past makes these collections less accessible. This specific form of contentiousness cannot be fully addressed by a tool built for detecting terms in the metadata.

Another limitation is that, although we adopted an inclusive approach to co-create the vocabulary with marginalized communities, we could not apply this process to the entire vocabulary. Our decision to focus the community work on specific themes and languages meant that we were unable to validate and discuss the vocabulary terms for all themes across every language. However, this was the only feasible option due to time and budget constraints. A full validation across all languages and by all affected communities would be a lengthy process, requiring substantial resources.

Finally, a third limitation pertains to the technical approach. The tool we provide as open source has high computational requirements, necessitating at least a mid-range GPU, which may limit accessibility for some who lack access to such resources. To address this, we make the API endpoint and the web application publicly available for anyone to use, with the goal of increasing accessibility. On the other hand, computational constraints prevented us from experimenting with larger models that could have potentially enhanced the tool’s performance. However, we do not view this as a significant limitation, as our model has demonstrated strong performance even with limited resources. By making the project open source, we enable teams with the necessary resources to build upon it and take it to the next level.

Ethical Considerations

This work was conducted in the context of an EU-funded initiative and adhered to the ethical standards set by the European Commission. As part of the funding requirements, the project underwent an independent ethics review prior to the start of the study, which concluded that the research was

in line with ethics and security requirements. An external advisory board of independent experts was appointed to provide ongoing oversight and guidance, particularly in relation to the use of potentially sensitive or harmful language. Their recommendations were regularly sought and integrated into the research process. All activities were also carried out in accordance with the institutional ethical guidelines of the partner organizations involved.

To support contributors involved in the vocabulary curation, community engagement, and detection evaluation activities, all participants were informed in advance about the sensitive nature of the content and were provided with an introductory session, inclusive guidelines, and content warnings. They were free to withdraw from any activity at any time without consequence. Remuneration and support for the community engagement were discussed and agreed upon with each group, reflecting local norms and the preferences of the participating communities. In some cases, such as in Belgium and the Netherlands, individual participants were directly compensated for their time and input. In others, including the UK and Germany, the available budget was directed toward the community more broadly—either by employing a community member over a longer period or compensating a workshop organizer from within the community. Indicative numbers include €200-€300 for individual participants and €750-€800 for workshop organizers, or €4,320 for a six-month community liaison role. These figures reflect what was feasible within the constraints of our budget and were always stipulated in mutual agreement with contributors. We believe that fair compensation is an important principle in collaborative research, especially when working with communities on sensitive topics. At the same time, we recognise that not all works will have the same resources. Transparency with contributors about available financial support is essential, and we provide these amounts for reference rather than as prescriptive standards for future work.

There are also certain ethical considerations regarding the technical developments and the use of resources created in this work. In our vocabulary, we provide suggested alternative terms for some vocabulary entries. When discussing our work with the CH community, a key concern was avoiding the simple replacement of contentious terms or “sanitizing” the metadata. While this was not our ap-

proach, we discussed the benefits of augmenting or enriching descriptions with those alternatives as this could offer the possibility to find CH objects when searching for these more appropriate alternative terms, improving accessibility. However, this approach was not pursued by us. It would require additional research since the applicability of the suggested terms depends on context, much like the contentiousness of many terms.

Another important ethical consideration is the potential harm of deploying an underprepared or low-performing harmful language detection tool. Flooding a platform with false positives could significantly erode user trust in the system, as users may begin to question its reliability and accuracy. Moreover, false flagging may unintentionally diminish the significance of the flagged issues in users’ minds. When contentious terms are flagged without proper context or accuracy, it risks trivializing important conversations and reducing awareness of these issues. Therefore, ensuring the tool is well-calibrated and accurate is crucial to maintaining both its credibility and its effectiveness in addressing harmful language.

Acknowledgements

This work has been partially supported by the project Detecting and Cur(at)ing Harmful Language in Cultural Heritage Collections (DE-BIAS), Grant No. 101100744, under the EU programme DIGITAL-2022-CULTURAL-02. It would not have been possible without the collaborative effort and dedication of all project partners. We are also grateful to the individuals and community members who participated in workshops, curated vocabulary terms, organized events, provided invaluable feedback, tested the tool, and supported this work in numerous ways. Their time, insights, trust and commitment to fostering more inclusive and reflective cultural heritage collections have been instrumental. The authors also gratefully acknowledge financial support from the Research Committee of the National Technical University of Athens (NTUA) through a scholarship.

References

- Pedro Alonso, Rajkumar Saini, and György Kovács. 2020. Hate speech detection using transformer ensembles on the hasoc dataset. In *International conference on speech and computer*, pages 13–21. Springer.

- Thomas Baker, Sean Bechhofer, Antoine Isaac, Alistair Miles, Guus Schreiber, and Ed Summers. 2013. Key choices in the design of simple knowledge organization system (skos). *Journal of Web Semantics*, 20:35–49.
- Michele Banko, Brendon MacKeen, and Laurie Ray. 2020. A unified taxonomy of harmful content. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 125–137, Online. Association for Computational Linguistics.
- Marco Bellagente, Jonathan Tow, Dakota Mahan, Duy Phung, Maksym Zhuravinskiy, Reshinh Adithyan, James Baicoianu, Ben Brooks, Nathan Cooper, Ashish Datta, et al. 2024. Stable lm 2 1.6 b technical report. *arXiv preprint arXiv:2402.17834*.
- Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Recent trends in word sense disambiguation: A survey. In *International Joint Conference on Artificial Intelligence*, pages 4330–4338. International Joint Conference on Artificial Intelligence, Inc.
- Meghana Moorthy Bhat, Saghar Hosseini, Ahmed Hassan, Paul Bennett, and Weisheng Li. 2021. Say ‘yes’ to positivity: Detecting toxic language in workplace communications. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2017–2029.
- Tom Bourgeade, Zongmin Li, Farah Benamara, Véronique Moriceau, Jian Su, and Aixun Sun. 2024. Humans need context, what about machines? investigating conversational context in abusive language detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8438–8452, Torino, Italia. ELRA and ICCL.
- Peter Burnap and Matthew Leighton Williams. 2014. Hate speech, machine classification and statistical modelling of information flows on twitter: Interpretation and communication for policy decision making.
- Tuhin Chakrabarty, Kilol Gupta, and Smaranda Muresan. 2019. Pay “attention” to your context when classifying abusive language. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 70–79, Florence, Italy. Association for Computational Linguistics.
- Kimberlé Williams Crenshaw. 2013. Mapping the margins: Intersectionality, identity politics, and violence against women of color. In *The public nature of private violence*, pages 93–118. Routledge.
- Sam Davidson, Qiusi Sun, and Magdalena Wojcieszak. 2020. Developing a new classifier for automated identification of incivility in social media. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 95–101, Online. Association for Computational Linguistics.
- Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.
- DE-BIAS Project. 2024. The DE-BIAS vocabulary and knowledge graph - documentation. https://pro.europeana.eu/files/Europeana_Professional/Projects/de_bias_vocabulary/TechnicalDocumentation_DE-BIAS_VocabularyKnowledgeGraph.pdf. Accessed: 2025-05-30.
- Fabio Del Vigna, Andrea Cimino, Felice Dell’Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the first Italian conference on cybersecurity (ITASEC17)*, pages 86–95.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- AbdelRahim Elmadany, Chiyu Zhang, Muhammad Abdul-Mageed, and Azadeh Hashemi. 2020. Leveraging affective bidirectional transformers for offensive language detection. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 102–108, Marseille, France. European Language Resource Association.
- European Commission, Directorate-General for Justice, Consumers, and L Kaati. 2023. *The European online hate lab*. Publications Office of the European Union.
- Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.
- Eirini Kaldeli, Orfeas Menis-Mastromichalakis, Spyros Bekiaris, Maria Ralli, Vassilis Tzouvaras, and Giorgos Stamou. 2021. Crowdheritage: crowdsourcing for improving the quality of cultural heritage metadata. *Information*, 12(02):64.
- Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.
- Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing hate speech classifiers with post-hoc explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5435–5442, Online. Association for Computational Linguistics.

- Katerina Korre, John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, Ion Androutsopoulos, Lucas Dixon, and Alberto Barrón-Cedeño. 2023. Harmful language datasets: An assessment of robustness. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 221–230.
- Jana Kurrek, Haji Mohammad Saleem, and Derek Ruths. 2020. [Towards a comprehensive taxonomy and large-scale annotated corpus for online slur usage](#). In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 138–149, Online. Association for Computational Linguistics.
- Daniel Loureiro, Kiamehr Rezaee, Mohammad Taher Pilehvar, and Jose Camacho-Collados. 2021. Analysis and evaluation of language models for word sense disambiguation. *Computational Linguistics*, 47(2):387–443.
- Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M Guerreiro, Ricardo Rei, Duarte M Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, et al. 2024. Eurollm: Multilingual language models for europe. *arXiv preprint arXiv:2409.16235*.
- Anneleen Masschelein, Frederik Truyen, Sofie Taes, Jonas Van Mulder, Andreas Stynen, and Roberta Pireddu. 2024. [Bias types and patterns: a typology applied to europeana use cases](#).
- Puneet Mathur, Rajiv Shah, Ramit Sawhney, and Debanjan Mahata. 2018. [Detecting offensive tweets in Hindi-English code-switched language](#). In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 18–26, Melbourne, Australia. Association for Computational Linguistics.
- Johannes Skjeggstad Meyer and Björn Gambäck. 2019. [A platform agnostic dual-strand hate speech detector](#). In *Proceedings of the Third Workshop on Abusive Language Online*, pages 146–156, Florence, Italy. Association for Computational Linguistics.
- Wayne Modest. 2018. *Words Matter: an unfinished guide to word choices in the cultural sector*. Tropenmuseum, Afrikamuseum, Museum Volkenkunde, Wereldmuseum.
- Sandip Modha, Prasenjit Majumder, and Thomas Mandl. 2018. [Filtering aggression from the multilingual social media feed](#). In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 199–207, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Andrei Nesterov, Laura Hollink, Marieke van Erp, and Jacco van Ossenbruggen. 2023. A knowledge graph of contentious terminology for inclusive representation of cultural heritage. In *European Semantic Web Conference*, pages 502–519. Springer.
- Nobal B. Niraula, Saurab Dulal, and Diwa Koirala. 2021. [Offensive language detection in Nepali social media](#). In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 67–75, Online. Association for Computational Linguistics.
- Flor Miriam Plaza-del Arco, Debora Nozza, Dirk Hovy, et al. 2023. Respectful or toxic? using zero-shot learning with language models to detect hate speech. In *The 7th Workshop on Online Abuse and Harms (WOAH)*. Association for Computational Linguistics.
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. [Stanza: A Python natural language processing toolkit for many human languages](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.
- Maria Ralli, Spyros Bekiaris, Eirini Kaldeli, Orfeas Menis-Mastromichalakis, Natasa Sofou, Vassilis Tzouvaras, and Giorgos Stamou. 2020. Crowdsourcing: Improving the quality of cultural heritage through crowdsourcing methods. In *2020 15th International Workshop on Semantic and Social Media Adaptation and Personalization (SMA)*, pages 1–6. IEEE.
- Haji Mohammad Saleem, Jana Kurrek, and Derek Ruths. 2022. Enriching abusive language detection with community context. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 131–142.
- Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. [The risk of racial bias in hate speech detection](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.
- A. Sasse. 2015. [Scaring and bullying people into security won't work](#). *IEEE Security & Privacy*, 13:80–83.
- Indira Sen, Mattia Samory, Claudia Wagner, and Isabelle Augenstein. 2022. [Counterfactually augmented data and unintended bias: The case of sexism and hate speech detection](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4716–4726, Seattle, United States. Association for Computational Linguistics.
- Hyeonho Song, Jisu Hong, Chani Jung, Hyojin Chin, Mingi Shin, Yubin Choi, Junghoi Choi, and Meeyoung Cha. 2024. Detecting offensive language in an open chatbot platform. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4760–4771.
- Melanie Subbiah, Amrita Bhattacharjee, Yilun Hua, Tharindu Kumarage, Huan Liu, and Kathleen MCKeown. 2023. [Towards detecting harmful agendas in](#)

- news articles. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 110–128, Toronto, Canada. Association for Computational Linguistics.
- TGDK Sumanathilaka, Nicholas Micallef, and Julian Hough. 2024. Can llms assist with ambiguity? a quantitative evaluation of various large language models on word sense disambiguation. *arXiv preprint arXiv:2411.18337*.
- Dias Oliva Thiago, Antonialli Dennys Marcelo, and Alessandra Gomes. 2021. Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to lgbtq voices online. *Sexuality & culture*, 25(2):700–732.
- Stéphan Tulkens, Lisa Hilte, Elise Lodewyckx, Ben Verhoeven, and Walter Daelemans. 2016. A dictionary-based approach to racism detection in dutch social media. *arXiv preprint arXiv:1608.08738*.
- William Warner and Julia Hirschberg. 2012. [Detecting hate speech on the world wide web](#). In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26, Montréal, Canada. Association for Computational Linguistics.
- Zeerak Waseem and Dirk Hovy. 2016. [Hateful symbols or hateful people? predictive features for hate speech detection on Twitter](#). In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.
- Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of abusive language: the problem of biased datasets. In *Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: human language technologies, volume 1 (long and short papers)*, pages 602–608.
- Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. 2020. [Demoting racial bias in hate speech detection](#). In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 7–14, Online. Association for Computational Linguistics.
- Jung H Yae, Nolan C Skelly, Neil C Ranly, and Phillip M LaCasse. 2024. Leveraging large language models for word sense disambiguation. *Neural Computing and Applications*, pages 1–18.
- Tsungcheng Yao, Ernest Foo, and Sebastian Binnewies. 2024. Personalised abusive language detection using llms and retrieval-augmented generation. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, pages 92–98.
- Guanhua Zhang, Bing Bai, Junqi Zhang, Kun Bai, Conghui Zhu, and Tiejun Zhao. 2020. [Demographics should not be the reason of toxicity: Mitigating discrimination in text classifications with instance weighting](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4134–4145, Online. Association for Computational Linguistics.
- Li Zhou, Antonia Karamolegkou, Wenyu Chen, and Daniel Hershcovich. 2023. [Cultural compass: Predicting transfer learning success in offensive language detection with cultural features](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12684–12702, Singapore. Association for Computational Linguistics.

A Vocabulary Examples

In Table 3 we present an example of a vocabulary entry.

B LLM sources and prompts

All LLMs used were found on Huggingface¹⁸. We used LLMs fine-tuned for instruction following, with the exception of the StableLM models which were only available as fine-tuned for chatting. We used model weights published in the GGUF format¹⁹ since that is compatible with llama.cpp. We provide URLs for the LLMs we used as footnotes: EuroLLM-9B²⁰ (Martins et al., 2024), Llama-3.1-8B²¹ (Dubey et al., 2024), Ministral-8B²², Mistral-Small-24B²³, Mixtral-8x7B²⁴ (Jiang et al., 2024), StableLM-2-1.6B²⁵ (Bellagente et al., 2024), StableLM-2-12B²⁶.

We produced a different prompt format for each language, starting with a base one in English and translating it to the other languages. In the preliminary stages, we tested using the same English prompt for all languages and simply denoting the language of the text, but this produced sub-par results. We provide as examples the English and Italian prompts we used for Llama-3.1-8B, as well as the English prompt for Mixtral8x7B in Table 4. The prompts used for different LLMs mostly

¹⁸<https://huggingface.co/>

¹⁹<https://huggingface.co/docs/hub/en/gguf>

²⁰<https://huggingface.co/bartowski/EuroLLM-9B-Instruct-GGUF>

²¹<https://huggingface.co/bartowski/Meta-Llama-3.1-8B-Instruct-GGUF>

²²<https://huggingface.co/bartowski/Ministral-8B-Instruct-2410-GGUF>

²³<https://huggingface.co/mistralai/Mistral-Small-24B-Instruct-2501>

²⁴<https://huggingface.co/TheBloke/Mixtral-8x7B-Instruct-v0.1-GGUF>

²⁵<https://huggingface.co/mradermacher/stablelm-2-1.6b-chat-i1-GGUF>

²⁶<https://huggingface.co/stabilityai/stablelm-2-12b-chat-GGUF>

Contentious Term	Contentious Issue	Suggestion Note	Suggested Term	Category
Caucasian	The term “Caucasian” was introduced in the late 18th century by German anthropologist Johann Friedrich Blumenbach, who idealized the Caucasus region as the origin of the “white race” and an aesthetic ideal. In the 19th century, it became a racial category encompassing Europeans, some North Africans, and parts of Western Asia, as part of pseudoscientific theories that hierarchically classified humans based on physical traits. Though later incorporated into Nazi racial ideology, the term was less central than narrower categories like “Aryan”. In the U.S., “Caucasian” became a common term for describing white populations in legal, demographic, and social contexts, often as a synonym for “white”. but it is increasingly criticized for its imprecision and roots in outdated racial theories	Use with caution in the context of people. The term is unproblematic when it refers specifically to people from the Caucasus region.	White	Ethnicity, Race

Table 3: A sample entry from the vocabulary.

differ on the special tokens that each LLM uses to separate system and user prompts.

One important aspect of the prompts is steering the output of the model towards an easily extractable answer. This was a problem we faced early on during our prompt engineering process, which we mitigated by a very structured format and by explicitly specifying the desired outputs. This very simplistic output format does not provide the model with enough tokens to “think” through the correct answer. We experimented with longer formats, including Chain-of-Thought, but they severely reduced the throughput at a level that was not acceptable for a tool meant to process millions of records. Nevertheless, in future versions of the tool we could provide longer prompt formats to users who do not prioritize throughput.

Model/Language	Prompt
Llama-3.1/English	<pre>< begin_of_text >< start_header_id >system< end_header_id > The term "{term}" can be contentious when used in some contexts. Here is an explanation of why "{term}" can be considered contentious: {vocabulary_context}< eot_id >< start_header_id >user< end_header_id >\nIs "{term}" used in a contentious manner in the following text? Answer with a simple yes or no.\n\nText:\n{text}< eot_id >< start_header_id >assistant < end_header_id ></pre>
Llama-3.1/Italian	<pre>< begin_of_text >< start_header_id >system< end_header_id > Il termine "{term}" può essere controverso quando utilizzato in alcuni contesti. Ecco una spiegazione del perché "{term}" può essere considerato controverso: {vocabulary_context} < eot_id >< start_header_id >user< end_header_id >\nIl termine "{term}" è utilizzato in modo controverso nel seguente testo? Rispondi con un semplice sì o no.\n\nTesto:\n{text}< eot_id >< start_header_id >assistant< end_header_id ></pre>
Mixtral8x7B/English	<pre><s> [INST] The term "{term}" can be contentious when used in some contexts. Here is an explanation of why "{term}" can be considered contentious: {vocabulary_context}\n\nQuestion:\nIs "{term}" used in a contentious manner in the following text? Answer with a simple yes or no.\n\nText:\n{text}\n\nAnswer:\n[/INST]\n</pre>

Table 4: Some of the prompts used for different LLMs and languages.