EHLD: A General Data Framework for Harmful Language Detection and Evaluation

Journal:	*Digital Library Perspectives*
Manuscript ID	DLP-01-2026-0028.R1
Manuscript Type:	Research Paper
Keywords:	Data Integration, LLMs, Data Evaluation, Hate Speech Detection, Harmful Language Detection, Italian Language

Digital Library Perspectives

EHLD: A General Data Framework for Harmful Language Detection and Evaluation

Purpose- Research into harmful language detection is hindered by fragmented datasets and incompatible label schemas, which significantly limit evaluation. This paper introduces EHLD: a general, extensible data framework supporting the integration, enrichment and evaluation of diverse harmful language resources, with a particular focus on Italian.

Design/Methodology/Approach- EHLD defines a unified schema with essential detection attributes (e.g., *harmful_or_not* and *harm_subcategory*), optional contextual dimensions (e.g., target, category, intersectionality), evaluation-oriented text complexity features, and metadata for provenance. We instantiate EHLD by integrating three data sources: curated datasets from the literature, large-scale LLM-based annotation from the *Mappa dell'Intolleranza*, and LLM generated data for inclusive language detection. We further introduce *label_confidence* to explicitly encode label reliability. Model selection follows an iterative cycle combining distributional analysis, linguistic diversity assessment, and performance evaluation.

Findings- The resulting dataset contains 237,956 instances with 18 features and exhibits substantial linguistic variety (Self-BLEU drops from 0.759 at 1-gram to 0.107 at 4-grams). After iterative, data-centric refinement, GPT-5-nano is selected as the reference model and achieves an F1-score of 0.813 on binary harmful language detection, outperforming reported baselines such as BERT-based multitask models and other LLMs on comparable Italian settings.

Research limitations/implications- Due to differences in datasets, domains, label definitions and evaluation protocols, comparisons across studies are not fully controlled. The framework partly relies on automatically generated labels, which, despite confidence-aware handling, may introduce residual noise.

Practical implications- By combining standardised labels, provenance tracking and complexity-oriented evaluation features, EHLD supports the construction of reproducible datasets and richer model auditing, enabling a more transparent deployment of harmful language detection systems.

Originality/value- EHLD contributes a reusable, confidence-aware integration framework that connects different harmful language resources and allows iterative, data-driven model selection and evaluation, with a focus on Italian.

Keywords: Harmful language Detection; Data Integration; Data Evaluation.

1. Introduction

Harmful language detection, in both monolingual and multilingual settings, has received increasing attention in recent years (D'Amico and Siccardi, 2020; Arora et al., 2023; Kiritchenko et al., 2021; Albladi et al., 2025), particularly with the development of large pre-trained and generative language models, which have substantially improved the ability to model complex linguistic patterns (Mahomed et al., 2024; Magnini et al., 2025). The term **harmful language** is an umbrella concept encom-

passing hate speech, stereotypes, offensive and abusive language, as well as other forms of linguistic bias that target, offend, or discriminate against individuals or social groups (Tamborini et al., 2025; Korre et al., 2023).

Despite this progress, research on harmful language detection remains highly fragmented (Arora et al., 2023). Existing datasets differ widely in terms of schema design, column definitions, annotation granularity, dataset size, and textual complexity (Poletto et al., 2021). Moreover, most datasets are constructed to address a single phenomenon, such as hate speech, toxicity, abuse, stereotypes, or harassment (Fortuna et al., 2020). They are typically annotated using incompatible label sets and are often limited to a single language, platform, or cultural context (Kumar et al., 2025).

This fragmentation has important implications for model development and evaluation. Models trained on individual datasets frequently achieve strong in-domain performance but exhibit limited generalization and robustness when applied to new platforms, languages, or target groups (Arango et al., 2019). Consequently, results reported across different studies are often difficult to compare, as they depend heavily on dataset-specific characteristics (Antypas and Camacho-Collados, 2023). This undermines the fairness and meaningfulness of evaluation practices and limits cumulative progress in the field.

In parallel, the increasing adoption of Large Language Models (LLMs) has intensified the demand for large, diverse, and reusable training resources, particularly for fine-tuning models for specific languages or downstream tasks (Zhang et al., 2025). However, the absence of a dedicated data integration infrastructure makes it challenging for researchers to effectively combine existing harmful language datasets. No standardized system currently provides access to multiple datasets through a single, shared schema.

Motivated by these limitations, our study proposes a general and extensible data framework for harmful language detection tasks. The primary objective of the framework is to provide a structured and reusable data schema that can be applied consistently across datasets, tasks, and evaluation settings. Based on a review of the literature, the framework defines a standardized set of core features that are essential for conducting harmful language detection tasks, while also identifying additional, non-mandatory contextual features that can enrich data representations and provide models with more informative task context. Furthermore, the framework introduces complementary features that are not directly related to the detection task itself, but that support more accurate and multifaceted model evaluation beyond reliance on a single label. We further demonstrate how data originating from heterogeneous sources can be systematically integrated into the proposed framework. The main contributions of this work can be summarized as follows:

- Aligning labels, values, and data structures to the extent possible, enabling dataset comparability and interoperability while reducing ad-hoc preprocessing.

- Improving scientific validity by supporting cross-dataset and cross-lingual learning, capturing shared structures of harmful language, and reducing dataset-specific biases.
- Enhancing reproducibility through standardized data splits, consistent evaluation metrics, and clearer benchmarking practices.
- Enabling modular and extensible research by facilitating the inclusion of new languages and phenomena, and by lowering the barrier for replication studies, low-resource language research, and longitudinal analyses.

The remainder of this paper is structured as follows. Section 2 reviews the background and related work in hate speech detection and dataset design. Section 3 presents the EHLD framework and provides the rationale for its architectural design. Section 4 describes our instantiation of EHLD for constructing a comprehensive dataset on harmful language in Italian, detailing the data integration and quality assessment procedures. Section 5 concludes with a discussion of the achieved performance results, limitations, and directions for future research.

2. Background and Related Works

The detection of **harmful language** has become a central research topic in *natural language processing*, driven by the proliferation of abusive, hateful, and discriminatory content on online platforms (Arora et al., 2023; Kiritchenko et al., 2021; Albladi et al., 2025). To support supervised learning and benchmarking, numerous datasets have been proposed, differing in language, annotation schemes, and conceptualizations of harmfulness (Poletto et al., 2021; Korre et al., 2023). While these datasets have advanced the field, their heterogeneity in structure and labeling conventions poses challenges for dataset interoperability, model comparison, and comprehensive evaluation (Rauh et al., 2022).

One of the most prominent datasets for Italian harmful language detection is **HaSpeeDe** (Del Vigna12 et al., 2017), released in the context of the EVALITA shared task. HaSpeeDe focuses on hate speech expressed in social media, primarily Twitter and Facebook, and frames the task as a binary classification problem, distinguishing hateful from non-hateful content. This formulation has enabled the development and comparison of a wide range of machine learning models. However, the dataset provides limited contextual information beyond the binary label, offering no explicit annotation of the targeted group, the specific type of harm, or linguistic phenomena such as implicit abuse or irony. As a result, HaSpeeDe is well suited for coarse-grained detection but less effective for fine-grained analysis of harmful language behavior.

To address some of these limitations, subsequent work has explored specific pragmatic aspects of harmful language. **IronITA** (Bosco et al., 2018) extends the scope of harmful language detection by focusing on irony and sarcasm in Italian texts. By explicitly annotating ironic content, IronITA highlights the difficulty of identifying harmful intent when it is conveyed indirectly or through rhetorical speech. While

this dataset enriches the representation of harmful language by capturing pragmatic complexity, its annotation schema is highly task-specific. The dataset is primarily designed for irony detection and does not provide a unified structure that can be easily aligned with other hate speech or abusive language resources, limiting its reuse in broader harmful language detection settings.

HateCheck (Röttger et al., 2021) introduces a functional test suite designed to systematically probe the behavior of hate speech detection models. Rather than relying on naturally occurring social media data, HateCheck consists of synthetically generated examples that cover a wide range of hate-related phenomena, including explicit and implicit abuse, protected and non-protected target groups. Among the datasets considered, HateCheck provides one of the most comprehensive and explicitly defined annotation schema, enabling fine-grained analysis of specific linguistic and semantic properties and supporting detailed diagnostic evaluation of model behavior. However, the simplicity and controlled nature of its synthetic examples limit their suitability for effective model training. The sentences are intentionally constructed to isolate individual phenomena, which makes them valuable for evaluation but less representative of the complexity, variability, and noise found in real-world harmful language.

Taken together, *HaSpeeDe*, *IronITA*, and *HateCheck* exemplify three complementary but largely disconnected approaches to harmful language data collection: coarse-grained binary classification on real-world data, phenomenon-specific annotation focusing on pragmatic cues, and synthetically generated test cases with rich annotation for diagnostic evaluation. The lack of a shared data schema across these resources hinders their integration and limits the ability to perform consistent analyses across datasets, tasks, and evaluation settings.

More recently, **LLMs** have increasingly been employed both as generators of harmful content and as tools for its detection and mitigation. As surveyed by Zhang et al. (2025), LLMs, on one hand, can amplify harmful language through uncontrolled generation, while on the other hand, they enable new approaches to safety, moderation, and data annotation. This growing reliance on LLMs has further highlighted the need for well-structured, interpretable, and reusable data representations that support robust evaluation across diverse harmful language phenomena.

Only recent work has begun to explore LLMs in harmful language annotation for Italian. For example, Leonardelli et al. (2025) introduce **MuLTa-Telegram**, a dataset annotated for hate speech and target detection in Italian and Polish. While this resource provides more detailed labels than earlier binary datasets, the annotated portion of the data remains relatively small, comprising approximately 2,000 examples for Italian, and exhibits an imbalance across target categories. These limitations restrict its applicability for large-scale training and systematic evaluation, particularly in settings that require balanced representations of different forms of harmful content.

Overall, the increased use of LLMs for both harmful language generation and

mitigation, together with the limited scale and heterogeneity of existing Italian datasets, underscores the need for a unified data framework. Such a framework should enable the integration of data from diverse sources, including manually annotated corpora, LLM-annotated data, and synthetically generated examples, while supporting fine-grained analysis and reliable model evaluation.

3. EHLD: A General Data Framework for Harmful Language Detection and Evaluation

Motivated by the limitations observed in existing datasets and annotation schemes, we propose EHLD (*Evaluation of Harmful Language Detection*), a *general* and *extensible* data framework for harmful language detection. EHLD defines a set of essential features, together with optional contextual and evaluation-oriented attributes, that can be used to represent harmful language data in a consistent and reusable manner. We demonstrate how data originating from heterogeneous sources, such as datasets collected from existing literature, data annotated using LLMs, and synthetically generated examples, can be integrated within EHLD.

The primary objective of EHLD is not to exhaustively capture all possible manifestations of harmful language, but rather to provide a structured and general-purpose data schema that can be applied across datasets, tasks, and evaluation settings, while remaining flexible to future extensions. In particular, it pursues the following goals:

(1) Define a standardized set of core features that are essential for conducting harmful language detection tasks across different domains and data sources.
(2) Identify additional, non-mandatory contextual features that can enrich the representation of harmful content and provide models with more informative task context.
(3) Introduce complementary features that are not directly tied to the detection task itself, but that enable more accurate, transparent, and multifaceted model evaluation beyond the use of a single label.
(4) Support extensibility, allowing new attributes, harm definitions, or evaluation dimensions to be incorporated as research practices and societal norms evolve.

Figure 1 illustrates the main stages of EHLD, including the collection of heterogeneous data sources, their normalization into the framework schema, and the subsequent use of the structured data for training, evaluation, and analysis. By explicitly separating essential detection features from contextual and evaluation-oriented information, our framework enables consistent comparison across datasets while preserving dataset-specific characteristics.

3.1. *Framework Structure and Attribute Organization*

Based on the above-mentioned goals, the EHLD framework organizes the data attributes into four main groups of columns:

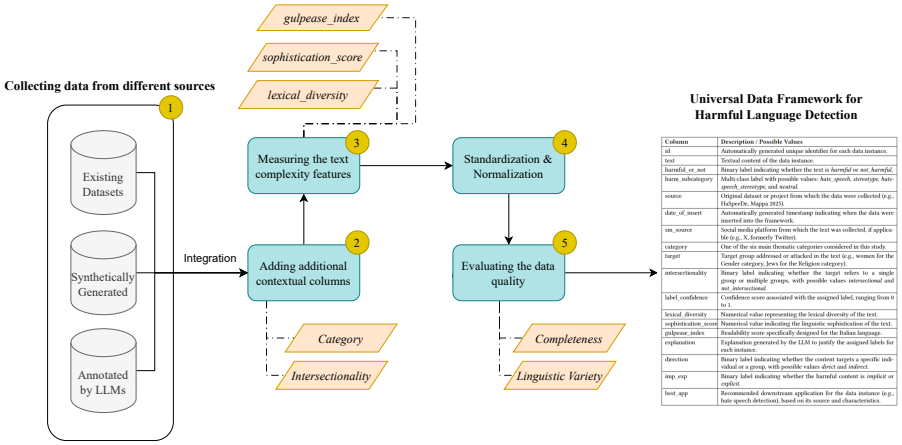

Fig. 1. The EHLD pipeline: main stages for integrating heterogeneous harmful language data sources into a unified framework schema.

- **Essential columns**, which include the minimum set of features required for model training and testing. The first column is "*id*", which is a unique identifier for each row of data, generated automatically using str(uuid.uuid4()) in Python. The second column is "text", which contains the main textual feature needed for analysis. The third column is a binary label called "*harmful_or_not*", indicating whether the text is harmful. This column helps us identify harmful texts in general. Besides this binary label, our definition of harmfulness includes various forms, such as hate speech, stereotypes, irony, inclusiveness, etc. After classifying the text as harmful or not using the previous column, we further specify the type of harm in the column "*harm_subcategory*".
- **Optional contextual columns**, which provide additional semantic information about the harmful content but may not be available in all datasets. The first column in this group is "*category*", which indicates the targeted attributes or identity markers. In our framework, we cover categories such as gender, sexual orientation, religion, ethnicity, and disability. After determining the category, we can identify the specific "*target*" group. For example, under the category gender, targets may be men or women, and under religion, targets may be Muslims or Jews. Sometimes a text targets multiple groups. In these cases, we use a binary column called "*intersectionality*", which indicates whether the text targets one entity or multiple entities simultaneously, e.g., Muslim women.
- **Evaluation-oriented columns**, which contain features designed to support a deeper analysis of model behavior and performance, for instance, by capturing properties related to text complexity.
- **Metadata columns**, including attributes such as the original data source or the date of ingestion into the framework, which preserve provenance information

and facilitate reproducibility.

Another contribution of this framework is the standardization of column values. Existing datasets often adopt heterogeneous representations for the same concepts—for example, using numerical labels (e.g., 0 and 1) in some datasets and textual labels (e.g., *hate* and *non-hate*) in others. After collecting data from three different sources to improve consistency and usability, we applied a series of standardization and normalization procedures to the values and columns included in the framework. Notably, the framework is designed to be flexible: while essential columns are required, optional and contextual columns may contain missing values when such information is unavailable.

Table 1 presents the complete list of columns defined within the framework, along with examples of the values associated with each attribute.

3.2. *Adding Contextual Information*

Annotation uncertainty is an inherent characteristic of harmful language datasets and can arise from multiple sources, including subjective label definitions, annotator disagreement, ambiguous linguistic expressions, and the use of automatic or semi-automatic annotation methods (Alaçam et al., 2025). As discussed in Section 2, recent work has increasingly explored the use of large language models (LLMs) to support data annotation; however, when annotations are not validated through a human-in-the-loop process involving domain experts, their correctness cannot be guaranteed with complete certainty.

Despite these limitations, prior research has shown that weakly supervised or noisy annotations, although not perfectly accurate, are often preferable to unlabeled data, particularly in large-scale learning settings where manual annotation is costly or infeasible (Zhou, 2018). Consequently, the proposed framework is designed to incorporate datasets with varying degrees of annotation reliability, regardless of whether the labels originate from human annotators, automatic systems, or hybrid approaches.

Several strategies can be adopted to estimate annotation uncertainty, including approaches based on inter-annotator agreement, model confidence scores, or the accuracy obtained on a set of well-curated and manually annotated data (Artstein, 2017). Rather than prescribing a single optimal solution, the proposed framework is designed to accommodate different uncertainty estimation methods, depending on the data source and annotation process.

To explicitly represent annotation uncertainty, the EHLD framework introduces a feature named *label_confidence*, which provides an approximate indication of the reliability of each label. When data originate from curated datasets in the literature and are annotated using human-in-the-loop approaches, such as expert annotation or crowdsourcing, the label confidence can be set to 1 to reflect the higher expected reliability of these annotations. Conversely, when labels are generated using automatic methods, such as LLM-based annotation, a simple heuristic can be adopted,

defining the confidence as the inverse of the total number of possible labels. For instance, if a model is required to select one label from four classes, the corresponding confidence score would be $1/4 = 0.25$. While this heuristic does not aim to provide a precise probabilistic estimate, encoding label confidence in this explicit manner enables downstream models and evaluation procedures to take annotation uncertainty into account during training and analysis.

In addition to uncertainty-related information, the framework introduces a *best_application* feature to improve data usability. Harmful language encompasses a range of related phenomena, including hate speech, abusive or offensive language, inclusiveness, stereotypes, and social bias. As a result, not all data instances are equally suitable for every downstream task. Based on the data source and the available annotations, this feature identifies the most appropriate application domain for each instance. This enables users to select task-relevant subsets of data more effectively and reduces the risk of mismatches between data characteristics and experimental objectives.

Finally, the framework includes a set of metadata and usability-oriented features that provide additional contextual information about each data instance. These features are summarized as follows:

(1) *label_confidence*: a value between 0 and 1 indicating the estimated reliability of the annotation.
(2) *best_application*: the downstream task for which the data instance is most appropriate.
(3) *source*: the original dataset or collection from which the data were obtained.
(4) *sm_source*: the type of platform or medium from which the data originate, such as social media platforms (e.g., X, Facebook).

3.3. *Adding Text Complexity Features*

As a key methodological contribution, we extend the proposed framework by introducing evaluation-oriented features that are largely absent from existing harmful language datasets. In many current approaches, model evaluation is primarily based on performance metrics, such as *accuracy* or F_1 score. These metrics quantify predictive performance but ignore the characteristics and quality of the underlying data. As a result, they cannot reveal the conditions under which models succeed or fail (Gong et al., 2023).

In particular, models may achieve high performance on test sets dominated by short, explicit, or lexically simple texts, while exhibiting substantially different behavior when applied to longer, more implicit, or linguistically complex inputs. Without explicitly characterizing these properties, it becomes difficult to determine whether observed performance reflects genuine robustness or is instead driven by biases in the dataset composition.

To address this limitation, the EHLD framework incorporates three metrics: *lexical_diversity*, *sophistication_score*, and *gulpease_index*. To measure these metrics,

we first convert all words to lowercase and then tokenize the text into sentences. To quantify lexical diversity, we use the *Type–Token Ratio (TTR)*, defined as $TTR = \frac{n}{N}$, wiht n the number of unique words and N the total number of words. This metric yields a value between 0 and 1: values close to 1 indicate a highly varied vocabulary, whereas values close to 0 indicate a high level of repetition.

The second metric, referred to as `sophistication_score`, is a simple proxy used to calculate the ratio of "less common" words to the total number of words. The list of common Italian words is defined as follows:

```
common_words_it = set([
    "il", "la", "e", "di", "che", "a", "in", "per", "un", "una",
    "con", "non", "su", "come", "da", "al", "del", "si", "mi",
    "ti", "lo", "gli", "le", "si"
])
```

For this metric, a higher score indicates a more content-heavy, informative, or technical text, while a lower score suggests a more conversational or simpler style.

Finally, the last metric is the `gulpease_index`, a readability formula created by the Gruppo Universitario Linguistico Pedagogico (GULP) at Rome's "La Sapienza" University to accurately measure the complexity of Italian texts (Lucisano and Piemontese, 1988). Unlike other readability indices, this formula relies on word length (measured in characters) and sentence length rather than syllable counts. The index ranges from 0 to 100, where higher values correspond to greater readability. The formula for it is as follows:

$$89 - 10 \cdot \frac{\text{letters}}{\text{words}} + 3 \cdot \frac{\text{sentences}}{\text{words}}$$

For interpretation, the index can be divided into four ranges:

- < 40: Very difficult (suitable for high-level education or university).
- < 60: Difficult (below 8th-grade level).
- < 80: Relatively easy (below 5th-grade level).
- $80 - 100$: Very easy and highly readable.

4. Building the EHLD Framework

Having presented the EHLD framework, we now introduce a concrete instantiation that systematically integrates three distinct data sources in order to construct a comprehensive, analysis-ready dataset. Through rigorous cleaning and normalisation procedures, we have developed a dataset characterised by informative, well-structured features. These features enable a thorough evaluation of trained models and provide valuable insights into the dataset's applicability for specific downstream tasks. More specifically, to validate the dataset's quality and demonstrate its analytical value, we have conducted comprehensive analyses of the data distributions

Table 1. Overview of the columns and their possible values in the EHLD framework

Column	**Description / Possible Values**
id	Automatically generated unique identifier for each data instance.
text	Textual content of the data instance.
harmful_or_not	Binary label indicating whether the text is *harmful* or *not_harmful*.
harm_subcategory	Multi-class label with possible values: *hate_speech*, *stereotype*, *hatespeech_stereotype*, and *neutral*.
source	Original dataset or project from which the data were collected (e.g., HaSpeeDe, Mappa 2025).
date_of_insert	Automatically generated timestamp indicating when the data were inserted into the framework.
sm_source	Social media platform from which the text was collected, if applicable (e.g., X, formerly Twitter).
category	One of the six main thematic categories considered in this study.
target	Target group addressed or attacked in the text (e.g., women for the Gender category, Jews for the Religion category).
intersectionality	Binary label indicating whether the target refers to a single group or multiple groups, with possible values *intersectional* and *not_intersectional*.
label_confidence	Confidence score associated with the assigned label, ranging from 0 to 1.
lexical_diversity	Numerical value representing the lexical diversity of the text.
sophistication_score	Numerical value indicating the linguistic sophistication of the text.
gulpease_index	Readability score specifically designed for the Italian language.
explanation	Explanation generated by the LLM to justify the assigned labels for each instance.
direction	Binary label indicating whether the content targets a specific individual or a group, with possible values *direct* and *indirect*.
imp_exp	Binary label indicating whether the harmful content is *implicit* or *explicit*.
best_app	Recommended downstream application for the data instance (e.g., hate speech detection), based on its source and characteristics.

within the framework and performed systematic assessments across multiple quality dimensions. Finally, we show that models trained on this dataset achieve accuracy results that surpass those previously reported for state-of-the-art performance.

4.1. *Data sources*

The data integrated into the proposed framework originate from three complementary sources: curated datasets from existing literature, large-scale annotation performed using LLMs, and synthetically generated data. This multi-source approach enables comprehensive coverage of established benchmarks while ensuring scalability for under-resourced tasks.

Existing literature. We incorporated three publicly available Italian datasets selected for their relevance and widespread adoption in prior research. **HaSpeeDe** (Del Vigna12 et al., 2017) provides binary hate speech annotations for social media texts from Twitter and Facebook. **IronITA** (Bosco et al., 2018) focuses on irony and sarcasm detection in Italian social media. **HateCheck** (Röttger et al., 2021) consists of functionally designed test cases that probe model robustness under diverse linguistic conditions. Collectively, these three datasets account for 6.58% of the total data included in the framework.

LLM-based annotation. The majority of our data (89.1%) originates from the *Map of Intolerance (Mappa dell'Intolleranza*[a]*)*, a national Italian initiative designed to identify discriminatory language on social media platforms (D'Amico and Siccardi, 2020; Fiano, 2025). In total, we collected and annotated 212,013 tweets across six distinct intolerance categories.

Synthetic label generation. To address data scarcity in under-explored tasks, we leverage LLMs to generate labels for inclusive language detection. Unlike traditional hate speech detection, inclusive language detection focuses on identifying subtle forms of exclusion, bias or non-inclusive wording rather than explicit, abusive or hostile expressions (D'Amico et al., 2023). Following a controlled labeling protocol with quality safeguards (Romano et al., 2024), we annotated 8,514 instances as either inclusive or non-inclusive. The provenance of each labeled instance is explicitly tracked within the framework metadata, enabling transparent evaluation and supporting analyses of performance differences across data sources. When combined with the original seed data, these LLM-labeled instances account for 3.58% of the total framework data.

4.2. *Data Cleaning and Normalization*

Another contribution of this framework is the systematic cleaning, integration, and standardization of heterogeneous datasets. Data collected from different sources often contain inconsistencies, noise, and divergent conventions that can negatively

[a]Visit `https://www.voxdiritti.it/ottava-edizione-mappa-intolleranza/` for a presentation of the results.

affect both model performance and downstream analyses. Typical issues include duplicated entries, inconsistent label formats, missing or ill-defined annotations, and variations in text encoding. Addressing these discrepancies is therefore a necessary prerequisite for building a reliable and reusable dataset.

Normalization is also crucial. Existing datasets frequently adopt heterogeneous representations for the same concepts, for example, using numerical labels (e.g., 0 and 1) in some datasets and textual labels (e.g., *hate* and *non-hate*) in others. To reconcile these differences, we applied a series of standardization, and normalization procedures after collecting data from three distinct sources. Labels originally expressed with different formats were mapped to a shared schema consisting of a binary *harmful / not harmful* label, indicating whether a text contains harmful content, and a multi-class *harm subcategory* label with possible values *hate speech*, *stereotype*, *hate speech + stereotype*, and *neutral*. This design allows us to harmonize data across sources while preserving a sufficient level of granularity to support fine-grained analyses of harmful content.

In addition, we assign a *confidence level* to each label to reflect how it was obtained, distinguishing between labels derived from manual annotation, weak supervision, or automated generation. This metadata enabled us a more nuanced analyses of model performance and data quality, and allows researchers to account for varying degrees of label reliability when training or evaluating models.

4.3. *Distribution of the data based on different features*

The resulting framework consists of 237,956 instances described by 18 features, whose categorization and intended purpose for each column are detailed Table 1. In this section, we analyze the distribution of the main features within the dataset and evaluate key data properties, including completeness and linguistic variety.

Table 2 presents the distribution of the two primary annotation dimensions: (i) the binary harmfulness label, which indicates whether a text contains harmful content, and (ii) the harmful language subcategory, which specifies the type of harmful content (e.g., hate speech, stereotype, etc.). The binary distribution shows an approximate 72% / 27% split between harmful and not-harmful instances. This imbalance is expected, as the framework intentionally covers multiple forms of harmful language, requiring a larger number of harmful examples to adequately capture their diversity. Regarding subcategories, nearly half of the harmful instances are labeled as hate speech, which can be attributed to the greater availability of publicly accessible resources and annotated datasets for this category compared to other forms of harmful language.

Table 3 illustrates the *subcategories* we have in the dataset, which represents the social attribute of the group under attack. Within our framework, approximately 45% of the instances belong to the **gender** category. The predominance of this category can be explained by two main factors. First, the existing literature shows a strong focus on gender-based abuse, with many studies addressing **misogyny**

Table 2. Harmful language and its subcategories distribution in our universal data framework.

harmful_or_not	**Distribution (%)**	**harm_subcategory**	**Distribution (%)**
harmful	72.14	hate_speech	50.47
		hatespeech_stereotype	10.54
		stereotype	9.29
		non_inclusive	0.98
		irony	0.47
		irony_sarcasm	0.38
not_harmful	27.86	neutral	25.27
		inclusive	2.59

as a primary case study (Frenda et al., 2019). Second, the synthetically generated portion of our dataset related to inclusive language places particular emphasis on gender, due to the inherently gendered nature of the Italian language.

For each category, we also report a subset of the most frequently occurring target groups, which helps clarify how targets are represented in the data and illustrates the meaning of the intersectionality feature. Overall, 86.93% of the data is non-intersectional, meaning that the text targets a single group, while 13.07% is intersectional, where multiple social groups are targeted simultaneously.

Table 3. Distribution of Target Groups by Category and Intersectionality

Category	**Distribution (%)**	**Target Group(s)**	**Share (%)**	**Intersectionality**
Gender	52.44	Women only	35.71	Not intersectional
		Women + Immigrants	4.84	Intersectional
		Men only	0.84	Not intersectional
		Women + LGBTQ+	0.65	Intersectional
		Women + Muslims	0.32	Intersectional
Religion	29.68	Jews only	23.48	Not intersectional
		Muslims + Jews	5.47	Intersectional
		Muslims + Immigrants	2.82	Intersectional
		Muslims only	2.66	Not intersectional
		Jews + Immigrants	1.15	Intersectional
		Muslims + Jews + Immigrants	0.73	Intersectional
		Muslims + Women	0.38	Intersectional
		Muslims + Immigrants + Jews*	0.38	Intersectional
Ethnicity	11.47	Immigrants only	9.77	Not intersectional
		Immigrants + Women	0.36	Intersectional
		Immigrants + Muslims	0.32	Intersectional
Disability	4.11	People with disabilities	4.3	Not intersectional
Sexual Orientation	2.3	LGBTQ+ only	2.41	Not intersectional
		LGBTQ+ + Women	0.23	Intersectional

**Note: This appears duplicate but represent different combinations in original data.*

4.4. *Text Complexity and Linguistic Variety*

As described in Section 3.3, three metrics were employed to quantify textual complexity within the proposed framework: the Gulpease Index, lexical diversity, and the sophistication score. Figures 2 to 5 present the average of these metrics across different dimensions of the framework, namely the original data source, harmful language subcategory, category, and the top ten most frequent target groups. Together, these analyses provide insight into how linguistic complexity varies across data characteristics and annotation dimensions.

4.4.1. *Text complexity by data source*

Figure 2 highlights the differences in text complexity across data sources. Curated datasets such as HaSpeeDe and IronITA exhibit higher Gulpease scores and lexical diversity, indicating more readable and lexically varied content, likely due to editorial filtering and annotation guidelines. In contrast, the synthetically generated data exhibit lower readability and reduced linguistic sophistication, reflecting a more uniform and constrained language structure. This effect is attributable to the masking-and-replacement strategy adopted during data generation, which limits lexical variability and promotes repetitive syntactic patterns by substituting specific spans of text with predefined alternatives.

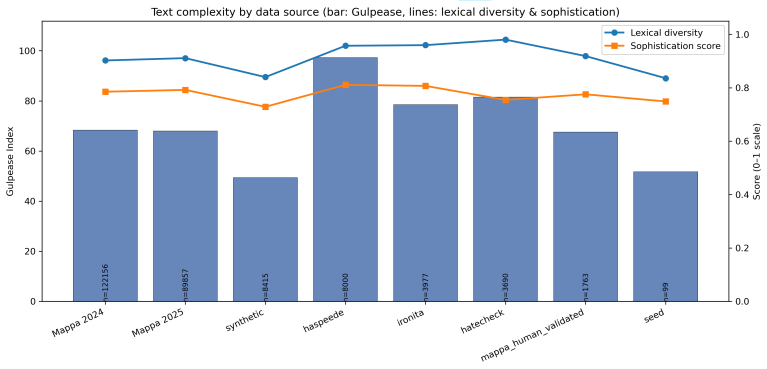

Fig. 2. Text complexity distribution by data source.

4.4.2. *Text complexity by harmful language subcategory*

Figure 3 shows that ironic and irony–sarcasm texts are characterized by higher readability and lexical diversity compared to direct hate speech or stereotype-related content. This suggests that indirect or figurative harmful expressions tend to employ richer and more varied language. Conversely, inclusive and non-inclusive categories

display lower Gulpease scores, indicating simpler and more explicit constructions, which is consistent with the synthetic-generated nature of such texts.

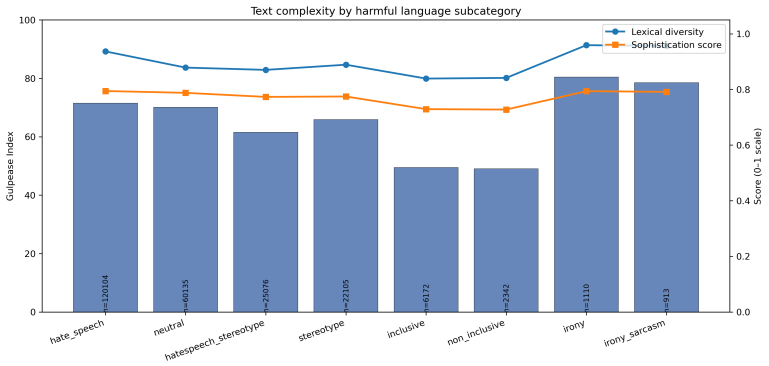

Fig. 3. Text complexity distribution across harmful language subcategories.

4.4.3. *Text complexity by target category*

As shown in Figure 4, texts targeting gender, ethnicity, and religion exhibit relatively similar levels of lexical diversity and sophistication. Overall, the relatively stable trends across categories suggest that text complexity is influenced more by discourse style than by the target category alone.

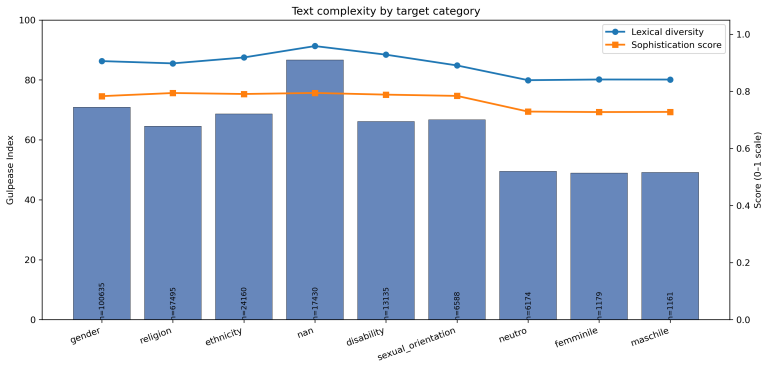

Fig. 4. Text complexity distribution by target category.

4.4.4. *Text complexity across target groups*

Figure 5 provides a more fine-grained view by examining the ten most frequent target groups, including single and intersectional targets. Mixed targets such as "jews + muslims" or "immigrants + women" tend to exhibit lower Gulpease scores and reduced lexical diversity, indicating more formulaic and repetitive language patterns. In contrast, single-target groups such as "women" or "disabled" show higher complexity values. This observation suggests that intersectional harmful content may rely on more stereotyped and less linguistically diverse expressions.

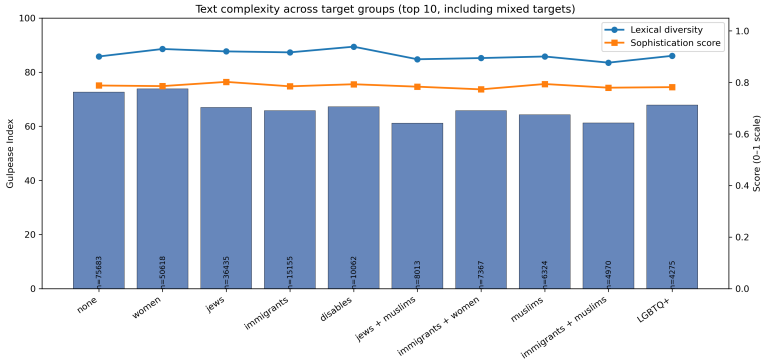

Fig. 5. Text complexity distribution across target groups.

4.4.5. *Linguistic Variety Assessment*

Another metric employed to assess the quality of the data collected in the framework is linguistic variety, measured using the Self-BLEU metric (Zhu et al., 2018). To compute this metric, we randomly sampled 1% of the overall dataset and calculated Self-BLEU scores for three different n-gram sizes, namely n=1, 2, and 4. The results indicate a high level of linguistic diversity within the dataset. Specifically, the Self-BLEU score decreases from 0.7591 at the unigram level, which reflects substantial lexical overlap across sentences, to 0.1073 at the 4-gram level, where similarity is measured over longer phrase sequences. This sharp decrease demonstrates that, while individual words may recur frequently, longer syntactic and semantic patterns are considerably more varied, suggesting a low level of redundancy and a high degree of textual diversity in the framework.

4.5. *Model Selection for Downstream Tasks*

Once the data have been collected, cleaned through quality checks, normalized, integrated into the framework, and enriched with additional features, they are used for model inference. This stage of the workflow aims to evaluate model performance

on the target downstream tasks and to iteratively improve results through controlled adaptations of both data and training strategies.

4.5.1. *Baseline Inference*

To establish a reference point for evaluation, we first perform baseline inference using publicly available open-source large language models obtained from repositories such as Hugging Face. These models are evaluated without any task-specific adaptation in order to measure their out-of-the-box performance. The resulting metrics serve as a benchmark against which models are compared and allow us to identify the most promising architectures for subsequent adaptation. Table 4 illustrates the results we achieved with our dataset.

Table 4. Baseline student models on binary harmful language detection

Model	**Acc**	**Macro**			**Weighted**		
		P	**R**	**F1**	**P**	**R**	**F1**
Gemma-3-4b-it	0.5831	0.7608	0.5637	0.4729	0.7517	0.5831	0.4837
Phi-4	0.6693	0.7799	0.6546	0.6209	0.7725	0.6693	0.6270
Ministral-3B	0.6168	0.7340	0.5910	0.5324	0.7240	0.6168	0.5457
Qwen3-4B	0.7157	0.7176	0.7125	0.7127	0.7171	0.7157	0.7140

4.5.2. *Dataset Formatting and Prompting*

Prompts are defined using a consistent template that provides task context, assigns a clear role to the model, and specifies the classification objective. Initial experiments are conducted using zero-shot prompts, while more advanced strategies, such as few-shot prompting and chain-of-thought–inspired formulations, are explored in later stages to improve output consistency and task understanding. Selected datasets are converted from their original tabular format into an instruction-following conversational structure compatible with the target models.

4.5.3. *Iterative Design and Evaluation Cycle*

Model selection and adaptation follows an iterative design lifecycle driven by integrated data-centric and model-centric evaluations. At each iteration, we systematically conduct:

- **distributional analysis** to identify class imbalance, source bias, and representation gaps,
- **linguistic diversity assessment** to ensure adequate coverage across lexical, syntactic, and stylistic variations,

- **performance evaluation** on the target downstream task using established metrics.

When performance outcomes doesn't meet the expectations, we do not rely solely on model slection and configuration. Instead, we implement targeted, data-level interventions such as rebalancing class distributions, adjusting the contribution weights of different data sources by changing the *label_confidence* values, and enriching linguistic diversity in underrepresented areas. This closed-loop process enables us to improve performance systematically while maintaining full transparency and rigorous control over data quality throughout the development cycle.

4.6. *Evaluation*

Model evaluation is conducted after completing the iterative design and evaluation cycle described in the previous section. Through successive rounds of data re balancing, linguistic enrichment, and prompt design, GPT-5-nano emerged as the most effective model in terms of both predictive performance and computational efficiency. As such, GPT-5-nano is selected as the reference model for the final evaluation and comparison with existing approaches.

Table 5 reports the per-class performance of GPT-5-nano on the binary harmful language detection task. The model achieves a high recall of 0.9283 for the `harmful` class, resulting in an F1-score of 0.8133. This indicates a strong capability to identify harmful content, which is a critical requirement in moderation-oriented scenarios. Conversely, the `not_harmful` class exhibits higher precision (0.8862) but lower recall (0.6119), reflecting a conservative bias toward minimizing false negatives for harmful content.

The selection of GPT-5-nano is the outcome of the iterative optimization process, in which performance improvements were achieved not only through model-centric tuning but also through targeted data-centric interventions. In particular, adjustments to data distributions, refinement of *label_confidence* weights across heterogeneous sources, and increased linguistic diversity in underrepresented patterns contributed to more stable and consistent performance across evaluation runs.

To contextualize these results, we compare GPT-5-nano with prior work on Italian harmful and hate speech detection, as summarized in Table 5. In the MULTA-TELEGRAM benchmark (Leonardelli et al., 2025), in-domain BERT-based multi-task models report an F1-score of 0.732 ± 0.025, while LLaMA-based models achieve F1-scores of 0.732. General-purpose safety models such as LLaMA-Guard achieve an F1-score of 0.712, highlighting the limitations of broad safety-oriented systems when applied to task-specific detection.

More recent large-scale models achieve higher absolute scores but at significantly higher computational cost. For example, (Magnini et al., 2025) report an F1-score of 0.77 for Qwen2.5-14B-Instruct-1M on hate speech detection, while top-performing systems in the HaSpeeDe 2 challenge at EVALITA (Del Vigna12 et al., 2017) achieve F1-macro scores of 0.80 on Twitter data and 0.77 on news headlines. These results

are obtained in competition settings using substantially larger models and task-specific training regimes.

Overall, the GPT-5-nano that we selected using our framework and adapted through prompt engineering outperforms the available baselines in the literature for the Italian language. This demonstrates the effectiveness of our proposed iterative, data-centric design methodology.

Table 5. Comparison of GPT-5-nano with prior work on Italian harmful/hate speech detection.

Model	Task	F1	Notes
Our approach			
			Selected using the EHLD framework
GPT-5-nano	Harmful language (binary)	**0.813**	Selected after iterative data-centric optimization
Related work			
			MuLTa-Telegram (Leonardelli et al., 2025)
BERT-based	Harmful language	0.732 ± 0.025	Manually annotated
LLaMA	Harmful language	0.732	General LLM baseline
LLaMA-Guard	Harmful language	0.712	General-purpose safety-focused model
			Benchmarking LLMs on Italian (Magnini et al., 2025)
Qwen2.5-14B	Hate speech	0.77	Instruction-tuned LLM
			HaSpeeDe 2 (Del Vigna12 et al., 2017)
BERT-based	Hate speech (Twitter)	0.80	Competition setting
BERT-based	Hate speech (Headlines)	0.77	Competition setting

5. Conclusion

This paper introduces **EHLD**, a general and extensible data framework designed to reduce fragmentation in harmful language detection research. It enables the integration of heterogeneous datasets into a unified, analysis-ready schema. EHLD combines essential detection labels with optional contextual attributes and **evaluation-oriented features** that capture text complexity. It also preserves provenance and supports reproducibility through its metadata. We instantiated the framework by integrating curated datasets from the literature, large-scale LLM-based annotation from the *Mappa dell'Intolleranza*, and LLM-labeled data for inclusive language detection, resulting in a consolidated dataset of 237,956 Italian instances described by

18 features.

A key outcome of the proposed methodology is the adoption of an **iterative, data-centric design** and evaluation cycle. In this cycle, model performance is improved through prompt design and model selection, as well as through systematic interventions on the data itself. Specifically, we analysed distributional properties to detect imbalances and gaps in representation, assessed **linguistic diversity** to identify underrepresented patterns, and revised the composition of the dataset by rebalancing and adjusting the contributions of the sources through *label_confidence*. Within this controlled loop, GPT-5-nano emerged as the most effective model under our constraints, offering an excellent balance of accuracy and efficiency.

In terms of performance, **GPT-5-nano** achieves an F1 score of 0.813 for binary *harmful* language detection, with a particularly high recall of 0.928 for the harmful class. While this behaviour is desirable in moderation-oriented settings where missing harmful content is risky, it is accompanied by lower recall for the *not_harmful* class, suggesting a conservative decision boundary. When contextualised against the results reported in the literature, the performance obtained exceeds the **baseline models** reported for Italian harmful language detection in comparable settings, such as BERT-based multitask approaches and other LLMs. Importantly, EHLD facilitates more interpretable comparisons by explicitly tracking provenance and label reliability, and by providing evaluation-oriented features that help explain when and why models succeed or fail.

Despite these contributions, several **limitations** remain. First, performance comparisons across studies are inherently constrained by differences in datasets, domains (e.g., tweets vs. headlines), annotation guidelines, and reported metrics. Second, although EHLD encodes label reliability via *label_confidence*, a substantial portion of the data originates from automatic annotation pipelines, which may contain residual noise that impacts both training and evaluation. Thirdly, while the schema is language-agnostic, the current instantiation is centred on Italian, and **cross-lingual validation** and multilingual integration have not yet been empirically demonstrated within this paper. Finally, the presented evaluation focuses on binary harmfulness, leaving robust multi-class and fine-grained target-level prediction as open challenges.

Future research can extend EHLD along two complementary axes. From a data perspective, richer coverage of underrepresented harmful phenomena (e.g., implicit abuse, coded language, and intersectional targeting) can be achieved through targeted data collection, expert validation, and the systematic enrichment of missing contextual attributes. From an evaluation perspective, EHLD can be expanded to support **fine-grained evaluation metrics** that go beyond aggregate scores, including detailed per-class precision, recall, and F1 scores. In addition, incorporating advanced deployment-oriented metrics such as the **abstention level** can better characterise model reliability under uncertainty and support risk-aware assessment in safety-critical moderation settings. Overall, EHLD provides a foundation for more

interoperable datasets and transparent evaluation practices, supporting cumulative progress towards robust systems for detecting harmful language.

Acknowledgment

The work reported in this paper has been partly funded by the European Union - NextGenerationEU, under the National Recovery and Resilience Plan (NRRP) Mission 4 Component 2 Investment Line 1.5: Strengthening of research structures and creation of R&D "innovation ecosystems", set up of "territorial leaders in R&D", within the project "MUSA - Multilayered Urban Sustainability Action" (contract no. ECS 00000037).

References

Alaçam, Ö., Hoeken, S., Säuberli, A., Gröner, H., Frassinelli, D., Zarrieß, S., and Plank, B. (2025). Disentangling subjectivity and uncertainty for hate speech annotation and modeling using gaze. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28695–28712.

Albladi, A., Islam, M., Das, A., Bigonah, M., Zhang, Z., Jamshidi, F., Rahgouy, M., Raychawdhary, N., Marghitu, D., and Seals, C. (2025). Hate speech detection using large language models: A comprehensive review. *IEEE Access*.

Antypas, D. and Camacho-Collados, J. (2023). Robust hate speech detection in social media: A cross-dataset empirical evaluation. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Arango, A., Pérez, J., and Poblete, B. (2019). Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, pages 45–54.

Arora, A., Nakov, P., Hardalov, M., Sarwar, S. M., Nayak, V., Dinkov, Y., Zlatkova, D., Dent, K., Bhatawdekar, A., Bouchard, G., et al. (2023). Detecting harmful content on online platforms: what platforms need vs. where research efforts go. *ACM Computing Surveys*, 56(3):1–17.

Artstein, R. (2017). Inter-annotator agreement. In *Handbook of linguistic annotation*, pages 297–313. Springer.

Bosco, C., Dell'Orletta, F., Poletto, F., Sanguinetti, M., Tesconi, M., et al. (2018). Overview of the evalita 2018 hate speech detection task. In *Ceur workshop proceedings*, volume 2263, pages 1–9. CEUR.

D'Amico, M. et al. (2023). *Parole che separano: linguaggio, Costituzione, diritti*, volume 151. Raffaello Cortina.

Del Vigna12, F., Cimino23, A., Dell'Orletta, F., Petrocchi, M., and Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the first Italian conference on cybersecurity (ITASEC17)*, pages 86–95.

D'Amico, M. and Siccardi, C. (2020). *VERSO UNA CULTURA CHE NON ODIA*. Giapichelli Editore, Milan, Italy.

Fiano, N. (2025). The "mappa dell'intolleranza" project. In *Book of Abstracts*, page 276.

Fortuna, P., Soler, J., and Wanner, L. (2020). Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6786–6794.

Frenda, S., Ghanem, B., Montes-y Gómez, M., and Rosso, P. (2019). Online hate speech against women: Automatic identification of misogyny and sexism on twitter. *Journal of intelligent & fuzzy systems*, 36(5):4743–4752.

Gong, Y., Liu, G., Xue, Y., Li, R., and Meng, L. (2023). A survey on dataset quality in machine learning. *Information and Software Technology*, 162:107268.

Kiritchenko, S., Nejadgholi, I., and Fraser, K. C. (2021). Confronting abusive language online: A survey from the ethical and human rights perspective. *Journal of Artificial Intelligence Research*, 71:431–478.

Korre, K., Pavlopoulos, J., Sorensen, J., Laugier, L., Androutsopoulos, I., Dixon, L., and Barrón-Cedeño, A. (2023). Harmful language datasets: An assessment of robustness. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 221–230.

Kumar, M. et al. (2025). Exploring hate speech detection: challenges, resources, current research and future directions. *Multimedia Tools and Applications*, pages 1–37.

Leonardelli, E., Casula, C., Salto, S. V., Bak, J. E., Muratore, E., Kołos, A., Louf, T., and Tonelli, S. (2025). Multa-telegram: A fine-grained italian and polish dataset for hate speech and target detection. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 582–594.

Lucisano, P. and Piemontese, M. (1988). Gulpease: una formula per la predizione della difficoltà dei testi in lingua italiana [gulpease: a formula to predict the difficulty of texts in italian language], in scuola e città, 3. *La Nuova Italia*.

Magnini, B., Madeddu, M., Resta, M., Zanoli, R., Cimmino, M., Albano, P., and Patti, V. (2025). A leaderboard for benchmarking llms on italian. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 636–646.

Mahomed, Y., Crawford, C. M., Gautam, S., Friedler, S. A., and Metaxa, D. (2024). Auditing gpt's content moderation guardrails: Can chatgpt write your favorite tv show? In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 660–686.

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., and Patti, V. (2021). Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55(2):477–523.

Rauh, M., Mellor, J., Uesato, J., Huang, P.-S., Welbl, J., Weidinger, L., Dathathri, S., Glaese, A., Irving, G., Gabriel, I., et al. (2022). Characteristics of harmful text: Towards rigorous benchmarking of language models. *Advances in Neural*

Information Processing Systems, 35:24720–24739.

Romano, T., Mohammadi, F., and Ceravolo, P. (2024). Synthetic data for identifying inclusive language (case study: Job descriptions in italian). In *2024 IEEE International Conference on Cyber Security and Resilience (CSR)*, pages 737–742. IEEE.

Röttger, P., Vidgen, B., Nguyen, D., Talat, Z., Margetts, H., and Pierrehumbert, J. (2021). Hatecheck: Functional tests for hate speech detection models. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pages 41–58.

Tamborini, M. A., Mohammadi, F., Ceravolo, P., Maghool, S., D'Amico, M. E., and Siccardi, C. (2025). Agent for anti-discriminatory language (aal) a human-centered ai approach to discriminatory speech in italian. In *2025 IEEE International Conference on Cyber Humanities (IEEE-CH)*, pages 1–6.

Zhang, C., Zhu, C., Xiong, J., Xu, X., Li, L., Liu, Y., and Lu, Z. (2025). Guardians and offenders: A survey on harmful content generation and safety mitigation of llm. *arXiv preprint arXiv:2508.05775*.

Zhou, Z.-H. (2018). A brief introduction to weakly supervised learning. *National science review*, 5(1):44–53.

Zhu, Y., Lu, S., Zheng, L., Guo, J., Zhang, W., Wang, J., and Yu, Y. (2018). Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1097–1100.

Biography

Digital Library Perspectives

EHLD: A General Data Framework for Harmful Language Detection and Evaluation

Purpose- Research into harmful language detection is hindered by fragmented datasets and incompatible label schemas, which significantly limit evaluation. This paper introduces EHLD: a general, extensible data framework supporting the integration, enrichment and evaluation of diverse harmful language resources, with a particular focus on Italian.
Design/Methodology/Approach- EHLD defines a unified schema with essential detection attributes (e.g., *harmful_or_not* and *harm_subcategory*), optional contextual dimensions (e.g., target, category, intersectionality), evaluation-oriented text complexity features, and metadata for provenance. We instantiate EHLD by integrating three data sources: curated datasets from the literature, large-scale LLM-based annotation from the *Mappa dell'Intolleranza*, and LLM generated data for inclusive language detection. We further introduce *label_confidence* to explicitly encode label reliability. Model selection follows an iterative cycle combining distributional analysis, linguistic diversity assessment, and performance evaluation.
Findings- The resulting dataset contains 237,956 instances with 18 features and exhibits substantial linguistic variety (Self-BLEU drops from 0.759 at 1-gram to 0.107 at 4-grams). After iterative, data-centric refinement, GPT-5-nano is selected as the reference model and achieves an F1-score of 0.813 on binary harmful language detection, outperforming reported baselines such as BERT-based multitask models and other LLMs on comparable Italian settings.
Research limitations/implications- Due to differences in datasets, domains, label definitions and evaluation protocols, comparisons across studies are not fully controlled. The framework partly relies on automatically generated labels, which, despite confidence-aware handling, may introduce residual noise.
Practical implications- By combining standardised labels, provenance tracking and complexity-oriented evaluation features, EHLD supports the construction of reproducible datasets and richer model auditing, enabling a more transparent deployment of harmful language detection systems.
Originality/value- EHLD contributes a reusable, confidence-aware integration framework that connects different harmful language resources and allows iterative, data-driven model selection and evaluation, with a focus on Italian.

Keywords: Harmful language Detection; Data Integration; Data Evaluation.

1. Introduction

Harmful language detection, in both monolingual and multilingual settings, has received increasing attention in recent years (D'Amico and Siccardi, 2020; Arora et al., 2023; Kiritchenko et al., 2021; Albladi et al., 2025), particularly with the development of large pre-trained and generative language models, which have substantially improved the ability to model complex linguistic patterns (Mahomed et al., 2024; Magnini et al., 2025). The term **harmful language** is an umbrella concept encom-

passing hate speech, stereotypes, offensive and abusive language, as well as other forms of linguistic bias that target, offend, or discriminate against individuals or social groups (Tamborini et al., 2025; Korre et al., 2023).

Despite this progress, research on harmful language detection remains highly fragmented (Arora et al., 2023). Existing datasets differ widely in terms of schema design, column definitions, annotation granularity, dataset size, and textual complexity (Poletto et al., 2021). Moreover, most datasets are constructed to address a single phenomenon, such as hate speech, toxicity, abuse, stereotypes, or harassment (Fortuna et al., 2020). They are typically annotated using incompatible label sets and are often limited to a single language, platform, or cultural context (Kumar et al., 2025).

This fragmentation has important implications for model development and evaluation. Models trained on individual datasets frequently achieve strong in-domain performance but exhibit limited generalization and robustness when applied to new platforms, languages, or target groups (Arango et al., 2019). Consequently, results reported across different studies are often difficult to compare, as they depend heavily on dataset-specific characteristics (Antypas and Camacho-Collados, 2023). This undermines the fairness and meaningfulness of evaluation practices and limits cumulative progress in the field.

In parallel, the increasing adoption of Large Language Models (LLMs) has intensified the demand for large, diverse, and reusable training resources, particularly for fine-tuning models for specific languages or downstream tasks (Zhang et al., 2025). This demand extends beyond NLP research: as LLMs are increasingly deployed in sensitive sociotechnical domains (Bellini, 2025), the availability of robust, structured, and interoperable data infrastructures becomes a broader societal concern. Yet despite this growing need, the harmful language detection field still lacks a dedicated data integration infrastructure, making it challenging for researchers to effectively combine existing datasets under a single, shared schema.

Motivated by these limitations, our study proposes a general and extensible data framework for harmful language detection tasks. The primary objective of the framework is to provide a structured and reusable data schema that can be applied consistently across datasets, tasks, and evaluation settings. Based on a review of the literature, the framework defines a standardized set of core features that are essential for conducting harmful language detection tasks, while also identifying additional, non-mandatory contextual features that can enrich data representations and provide models with more informative task context. Furthermore, the framework introduces complementary features that are not directly related to the detection task itself, but that support more accurate and multifaceted model evaluation beyond reliance on a single label. We further demonstrate how data originating from heterogeneous sources can be systematically integrated into the proposed framework. The main contributions of this work can be summarized as follows:

- Aligning labels, values, and data structures to the extent possible, enabling

dataset comparability and interoperability while reducing ad-hoc preprocessing.
- Improving scientific validity by supporting cross-dataset and cross-lingual learning, capturing shared structures of harmful language, and reducing dataset-specific biases.
- Enhancing reproducibility through standardized data splits, consistent evaluation metrics, and clearer benchmarking practices.
- Enabling modular and extensible research by facilitating the inclusion of new languages and phenomena, and by lowering the barrier for replication studies, low-resource language research, and longitudinal analyses.

The remainder of this paper is structured as follows. Section 2 reviews the background and related work in hate speech detection and dataset design. Section 3 presents the EHLD framework and provides the rationale for its architectural design. Section 4 describes our instantiation of EHLD for constructing a comprehensive dataset on harmful language in Italian, detailing the data integration and quality assessment procedures. Section 5 concludes with a discussion of the achieved performance results, limitations, and directions for future research.

2. Background and Related Works

The detection of **harmful language** has become a central research topic in *natural language processing*, driven by the proliferation of abusive, hateful, and discriminatory content on online platforms (Arora et al., 2023; Kiritchenko et al., 2021; Albladi et al., 2025). To support supervised learning and benchmarking, numerous datasets have been proposed, differing in language, annotation schemes, and conceptualizations of harmfulness (Poletto et al., 2021; Korre et al., 2023). While these datasets have advanced the field, their heterogeneity in structure and labeling conventions poses challenges for dataset interoperability, model comparison, and comprehensive evaluation (Rauh et al., 2022).

These challenges are further compounded in Italian, where harmful language is strongly shaped by grammatical gender, cultural norms, and indirect expression. Discriminatory content frequently manifests through irony, sarcasm, or culturally specific idioms rather than explicit slurs Frenda (2023), and the language's pervasive grammatical gender has fuelled ongoing debates around bias and inclusivity Maghool et al. (2025); Baiocco et al. (2023). As a result, annotation schemes developed for English cannot be transferred to Italian without substantial adaptation (Poletto et al., 2021). This is compounded by a severe resource asymmetry: as of March 2026, Hugging Face lists 87,679 English text-generation models against only 5,129 for Italian (a 17-to-1 ratio), and 17,890 English datasets against 653 in Italian (27-to-1). No Italian-specific LLM comparable in scale to GPT-4 or Llama exists, and the available Italian models — AlBERTo Polignano et al. (2019) and UmBERTo Magnini et al. (2006) — are modest BERT-base systems trained on far smaller corpora. When applying LLMs to Italian harmful language detection, one

therefore relies on models with limited Italian exposure, likely contributing to erratic classification behaviour on nuanced phenomena such as irony and stereotypes.

One of the most prominent datasets for Italian harmful language detection is **HaSpeeDe** (Del Vigna12 et al., 2017), released in the context of the EVALITA shared task. HaSpeeDe focuses on hate speech expressed in social media, primarily Twitter and Facebook, and frames the task as a binary classification problem, distinguishing hateful from non-hateful content. This formulation has enabled the development and comparison of a wide range of machine learning models. However, the dataset provides limited contextual information beyond the binary label, offering no explicit annotation of the targeted group, the specific type of harm, or linguistic phenomena such as implicit abuse or irony. As a result, HaSpeeDe is well suited for coarse-grained detection but less effective for fine-grained analysis of harmful language behavior.

To address some of these limitations, subsequent work has explored specific pragmatic aspects of harmful language. **IronITA** (Bosco et al., 2018) extends the scope of harmful language detection by focusing on irony and sarcasm in Italian texts. By explicitly annotating ironic content, IronITA highlights the difficulty of identifying harmful intent when it is conveyed indirectly or through rhetorical speech. While this dataset enriches the representation of harmful language by capturing pragmatic complexity, its annotation schema is highly task-specific. The dataset is primarily designed for irony detection and does not provide a unified structure that can be easily aligned with other hate speech or abusive language resources, limiting its reuse in broader harmful language detection settings.

HateCheck (Röttger et al., 2021) introduces a functional test suite designed to systematically probe the behavior of hate speech detection models. Rather than relying on naturally occurring social media data, HateCheck consists of synthetically generated examples that cover a wide range of hate-related phenomena, including explicit and implicit abuse, protected and non-protected target groups. Among the datasets considered, HateCheck provides one of the most comprehensive and explicitly defined annotation schema, enabling fine-grained analysis of specific linguistic and semantic properties and supporting detailed diagnostic evaluation of model behavior. However, the simplicity and controlled nature of its synthetic examples limit their suitability for effective model training. The sentences are intentionally constructed to isolate individual phenomena, which makes them valuable for evaluation but less representative of the complexity, variability, and noise found in real-world harmful language.

Taken together, *HaSpeeDe*, *IronITA*, and *HateCheck* exemplify three complementary but largely disconnected approaches to harmful language data collection: coarse-grained binary classification on real-world data, phenomenon-specific annotation focusing on pragmatic cues, and synthetically generated test cases with rich annotation for diagnostic evaluation. The lack of a shared data schema across these resources hinders their integration and limits the ability to perform consistent anal-

yses across datasets, tasks, and evaluation settings.

More recently, **LLMs** have increasingly been employed both as generators of harmful content and as tools for its detection and mitigation. As surveyed by Zhang et al. (2025), LLMs, on one hand, can amplify harmful language through uncontrolled generation, while on the other hand, they enable new approaches to safety, moderation, and data annotation. This growing reliance on LLMs has further highlighted the need for well-structured, interpretable, and reusable data representations that support robust evaluation across diverse harmful language phenomena.

Only recent work has begun to explore LLMs in harmful language annotation for Italian. For example, Leonardelli et al. (2025) introduce **MuLTa-Telegram**, a dataset annotated for hate speech and target detection in Italian and Polish. While this resource provides more detailed labels than earlier binary datasets, the annotated portion of the data remains relatively small, comprising approximately 2,000 examples for Italian, and exhibits an imbalance across target categories. These limitations restrict its applicability for large-scale training and systematic evaluation, particularly in settings that require balanced representations of different forms of harmful content.

Overall, the increased use of LLMs for both harmful language generation and mitigation, together with the limited scale and heterogeneity of existing Italian datasets, underscores the need for a unified data framework. Such a framework should enable the integration of data from diverse sources, including manually annotated corpora, LLM-annotated data, and synthetically generated examples, while supporting fine-grained analysis and reliable model evaluation.

3. EHLD: A General Data Framework for Harmful Language Detection and Evaluation

Motivated by the limitations observed in existing datasets and annotation schemes, we propose EHLD (*Evaluation of Harmful Language Detection*), a *general* and *extensible* data framework for harmful language detection. EHLD defines a set of essential features, together with optional contextual and evaluation-oriented attributes, that can be used to represent harmful language data in a consistent and reusable manner. We demonstrate how data originating from heterogeneous sources, such as datasets collected from existing literature, data annotated using LLMs, and synthetically generated examples, can be integrated within EHLD.

The primary objective of EHLD is not to exhaustively capture all possible manifestations of harmful language, but rather to provide a structured and general-purpose data schema that can be applied across datasets, tasks, and evaluation settings, while remaining flexible to future extensions. In particular, it pursues the following goals:

(1) Define a standardized set of core features that are essential for conducting harmful language detection tasks across different domains and data sources.
(2) Identify additional, non-mandatory contextual features that can enrich the rep-

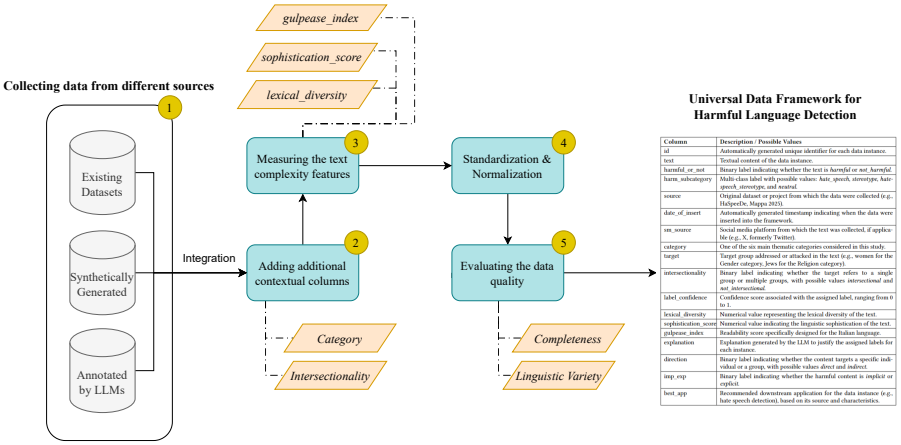

Fig. 1. The EHLD pipeline: main stages for integrating heterogeneous harmful language data sources into a unified framework schema.

resentation of harmful content and provide models with more informative task context.

(3) Introduce complementary features that are not directly tied to the detection task itself, but that enable more accurate, transparent, and multifaceted model evaluation beyond the use of a single label.

(4) Support extensibility, allowing new attributes, harm definitions, or evaluation dimensions to be incorporated as research practices and societal norms evolve.

Figure 1 illustrates the main stages of EHLD, including the collection of heterogeneous data sources, their normalization into the framework schema, and the subsequent use of the structured data for training, evaluation, and analysis. By explicitly separating essential detection features from contextual and evaluation-oriented information, our framework enables consistent comparison across datasets while preserving dataset-specific characteristics.

3.1. *Framework Structure and Attribute Organization*

Based on the above-mentioned goals, the EHLD framework organizes the data attributes into four main groups of columns:

- **Essential columns**, which include the minimum set of features required for model training and testing. The first column is "*id*", which is a unique identifier for each row of data, generated automatically using str(uuid.uuid4()) in Python. The second column is "text", which contains the main textual feature needed for analysis. The third column is a binary label called "*harmful_or_not*", indicating whether the text is harmful. This column helps us identify harmful texts in general. Besides this binary label, our definition of harmfulness includes

various forms, such as hate speech, stereotypes, irony, inclusiveness, etc. After classifying the text as harmful or not using the previous column, we further specify the type of harm in the column "*harm_subcategory*".

- **Optional contextual columns**, which provide additional semantic information about the harmful content but may not be available in all datasets. The first column in this group is "*category*", which indicates the targeted attributes or identity markers. In our framework, we cover categories such as gender, sexual orientation, religion, ethnicity, and disability. After determining the category, we can identify the specific "*target*" group. For example, under the category gender, targets may be men or women, and under religion, targets may be Muslims or Jews. Sometimes a text targets multiple groups. In these cases, we use a binary column called "*intersectionality*", which indicates whether the text targets one entity or multiple entities simultaneously, e.g., Muslim women.
- **Evaluation-oriented columns**, which contain features designed to support a deeper analysis of model behavior and performance, for instance, by capturing properties related to text complexity.
- **Metadata columns**, including attributes such as the original data source or the date of ingestion into the framework, which preserve provenance information and facilitate reproducibility.

Another contribution of this framework is the standardization of column values. Existing datasets often adopt heterogeneous representations for the same concepts—for example, using numerical labels (e.g., 0 and 1) in some datasets and textual labels (e.g., *hate* and *non-hate*) in others. After collecting data from three different sources to improve consistency and usability, we applied a series of standardization and normalization procedures to the values and columns included in the framework. Notably, the framework is designed to be flexible: while essential columns are required, optional and contextual columns may contain missing values when such information is unavailable.

Table 1 presents the complete list of columns defined within the framework, along with examples of the values associated with each attribute.

3.2. *Adding Contextual Information*

Annotation uncertainty is an inherent characteristic of harmful language datasets and can arise from multiple sources, including subjective label definitions, annotator disagreement, ambiguous linguistic expressions, and the use of automatic or semi-automatic annotation methods (Alaçam et al., 2025). As discussed in Section 2, recent work has increasingly explored the use of large language models (LLMs) to support data annotation; however, when annotations are not validated through a human-in-the-loop process involving domain experts, their correctness cannot be guaranteed with complete certainty.

Despite these limitations, prior research has shown that weakly supervised or noisy annotations, although not perfectly accurate, are often preferable to unlabeled

data, particularly in large-scale learning settings where manual annotation is costly or infeasible (Zhou, 2018). Consequently, the proposed framework is designed to incorporate datasets with varying degrees of annotation reliability, regardless of whether the labels originate from human annotators, automatic systems, or hybrid approaches.

Several strategies can be adopted to estimate annotation uncertainty, including approaches based on inter-annotator agreement, model confidence scores, or the accuracy obtained on a set of well-curated and manually annotated data (Artstein, 2017). Rather than prescribing a single optimal solution, the proposed framework is designed to accommodate different uncertainty estimation methods, depending on the data source and annotation process.

To explicitly represent annotation uncertainty, the EHLD framework introduces a feature named *label_confidence*, which provides an approximate measure of the reliability of each label. Labels derived from curated datasets and produced through human-in-the-loop approaches — such as expert annotation or crowdsourcing — can be assigned a confidence of 1, reflecting their higher expected reliability. In contrast, labels generated by automatic methods, such as LLM-based annotation, are assigned a confidence score using a simple heuristic: the inverse of the number of possible classes. For example, if a model must choose among four classes, the confidence score is $1/4 = 0.25$. Although this heuristic is not intended as a precise probabilistic estimate, its underlying rationale is to assign greater weight to human-in-the-loop annotations, which are generally considered more reliable than those produced automatically. By explicitly encoding this distinction, downstream models and evaluation procedures can prioritize higher-confidence labels during training and analysis, rather than treating all annotations as equally trustworthy.

In addition to uncertainty-related information, the framework introduces a *best_application* feature to improve data usability. Harmful language encompasses a range of related phenomena, including hate speech, abusive or offensive language, inclusiveness, stereotypes, and social bias. As a result, not all data instances are equally suitable for every downstream task. Based on the data source and the available annotations, this feature identifies the most appropriate application domain for each instance. This enables users to select task-relevant subsets of data more effectively and reduces the risk of mismatches between data characteristics and experimental objectives.

Finally, the framework includes a set of metadata and usability-oriented features that provide additional contextual information about each data instance. These features are summarized as follows:

(1) *label_confidence*: a value between 0 and 1 indicating the estimated reliability of the annotation.
(2) *best_application*: the downstream task for which the data instance is most appropriate.
(3) *source*: the original dataset or collection from which the data were obtained.

(4) *sm_source*: the type of platform or medium from which the data originate, such as social media platforms (e.g., X, Facebook).

3.3. *Adding Text Complexity Features*

As a key methodological contribution, we extend the proposed framework by introducing evaluation-oriented features that are largely absent from existing harmful language datasets. In many current approaches, model evaluation is primarily based on performance metrics, such as *accuracy* or F_1 score. These metrics quantify predictive performance but ignore the characteristics and quality of the underlying data. As a result, they cannot reveal the conditions under which models succeed or fail (Gong et al., 2023).

In particular, models may achieve high performance on test sets dominated by short, explicit, or lexically simple texts, while exhibiting substantially different behavior when applied to longer, more implicit, or linguistically complex inputs. Without explicitly characterizing these properties, it becomes difficult to determine whether observed performance reflects genuine robustness or is instead driven by biases in the dataset composition.

To address this limitation, the EHLD framework incorporates three metrics: *lexical_diversity*, *sophistication_score*, and *gulpease_index*. To measure these metrics, we first convert all words to lowercase and then tokenize the text into sentences. To quantify lexical diversity, we use the *Type–Token Ratio (TTR)*, defined as $TTR = \frac{n}{N}$, wiht n the number of unique words and N the total number of words. This metric yields a value between 0 and 1: values close to 1 indicate a highly varied vocabulary, whereas values close to 0 indicate a high level of repetition.

The second metric, referred to as `sophistication_score`, is a simple proxy used to calculate the ratio of "less common" words to the total number of words. The list of common Italian words is defined as follows:

```
common_words_it = set([
    "il", "la", "e", "di", "che", "a", "in", "per", "un", "una",
    "con", "non", "su", "come", "da", "al", "del", "si", "mi",
    "ti", "lo", "gli", "le", "si"
])
```

For this metric, a higher score indicates a more content-heavy, informative, or technical text, while a lower score suggests a more conversational or simpler style.

Finally, the last metric is the `gulpease_index`, a readability formula created by the Gruppo Universitario Linguistico Pedagogico (GULP) at Rome's "La Sapienza" University to accurately measure the complexity of Italian texts (Lucisano and Piemontese, 1988). Unlike other readability indices, this formula relies on word length (measured in characters) and sentence length rather than syllable counts. The index ranges from 0 to 100, where higher values correspond to greater readability. The formula for it is as follows:

$$89 - 10 \cdot \frac{\text{letters}}{\text{words}} + 3 \cdot \frac{\text{sentences}}{\text{words}}$$

For interpretation, the index can be divided into four ranges:

- < 40: Very difficult (suitable for high-level education or university).
- < 60: Difficult (below 8th-grade level).
- < 80: Relatively easy (below 5th-grade level).
- $80 - 100$: Very easy and highly readable.

4. Building the EHLD Framework

Having presented the EHLD framework, we now introduce a concrete instantiation that systematically integrates three distinct data sources in order to construct a comprehensive, analysis-ready dataset. Through rigorous cleaning and normalisation procedures, we have developed a dataset characterised by informative, well-structured features. These features enable a thorough evaluation of trained models and provide valuable insights into the dataset's applicability for specific downstream tasks. More specifically, to validate the dataset's quality and demonstrate its analytical value, we have conducted comprehensive analyses of the data distributions within the framework and performed systematic assessments across multiple quality dimensions. Finally, we show that models trained on this dataset achieve accuracy results that surpass those previously reported for state-of-the-art performance.

4.1. *Data sources*

The data integrated into the proposed framework originate from three complementary sources: curated datasets from existing literature, large-scale annotation performed using LLMs, and synthetically generated data. This multi-source approach enables comprehensive coverage of established benchmarks while ensuring scalability for under-resourced tasks.

Existing literature. We incorporated three publicly available Italian datasets selected for their relevance and widespread adoption in prior research. **HaSpeeDe** (Del Vigna12 et al., 2017) provides binary hate speech annotations for social media texts from Twitter and Facebook. **IronITA** (Bosco et al., 2018) focuses on irony and sarcasm detection in Italian social media. **HateCheck** (Röttger et al., 2021) consists of functionally designed test cases that probe model robustness under diverse linguistic conditions. Collectively, these three datasets account for 6.58% of the total data included in the framework.

LLM-based annotation. The majority of our data (89.1%) originates from the *Map of Intolerance (Mappa dell'Intolleranza*[a]*)*, a national Italian initiative de-

[a]Visit `https://www.voxdiritti.it/ottava-edizione-mappa-intolleranza/` for a presentation of the results.

Table 1. Overview of the columns and their possible values in the EHLD framework

Column	**Description / Possible Values**
id	Automatically generated unique identifier for each data instance.
text	Textual content of the data instance.
harmful_or_not	Binary label indicating whether the text is *harmful* or *not_harmful*.
harm_subcategory	Multi-class label with possible values: *hate_speech*, *stereotype*, *hatespeech_stereotype*, and *neutral*.
source	Original dataset or project from which the data were collected (e.g., HaSpeeDe, Mappa 2025).
date_of_insert	Automatically generated timestamp indicating when the data were inserted into the framework.
sm_source	Social media platform from which the text was collected, if applicable (e.g., X, formerly Twitter).
category	One of the six main thematic categories considered in this study.
target	Target group addressed or attacked in the text (e.g., women for the Gender category, Jews for the Religion category).
intersectionality	Binary label indicating whether the target refers to a single group or multiple groups, with possible values *intersectional* and *not_intersectional*.
label_confidence	Confidence score associated with the assigned label, ranging from 0 to 1.
lexical_diversity	Numerical value representing the lexical diversity of the text.
sophistication_score	Numerical value indicating the linguistic sophistication of the text.
gulpease_index	Readability score specifically designed for the Italian language.
explanation	Explanation generated by the LLM to justify the assigned labels for each instance.
direction	Binary label indicating whether the content targets a specific individual or a group, with possible values *direct* and *indirect*.
imp_exp	Binary label indicating whether the harmful content is *implicit* or *explicit*.
best_app	Recommended downstream application for the data instance (e.g., hate speech detection), based on its source and characteristics.

signed to identify discriminatory language on social media platforms (D'Amico and Siccardi, 2020; Fiano, 2025). In total, we collected and annotated 212,013 tweets across six distinct intolerance categories.

Synthetic label generation. To address data scarcity in under-explored tasks, we leverage LLMs to generate labels for inclusive language detection. Unlike traditional hate speech detection, inclusive language detection focuses on identifying subtle forms of exclusion, bias or non-inclusive wording rather than explicit, abusive or hostile expressions (D'Amico et al., 2023). Following a controlled labeling protocol with quality safeguards (Romano et al., 2024), we annotated 8,514 instances as either inclusive or non-inclusive. The provenance of each labeled instance is explicitly tracked within the framework metadata, enabling transparent evaluation and supporting analyses of performance differences across data sources. When combined with the original seed data, these LLM-labeled instances account for 3.58% of the total framework data.

4.2. *Data Cleaning and Normalization*

Another contribution of this framework is the systematic cleaning, integration, and standardization of heterogeneous datasets. Data collected from different sources often contain inconsistencies, noise, and divergent conventions that can negatively affect both model performance and downstream analyses. Typical issues include duplicated entries, inconsistent label formats, missing or ill-defined annotations, and variations in text encoding. Addressing these discrepancies is therefore a necessary prerequisite for building a reliable and reusable dataset.

Normalization is also crucial. Existing datasets frequently adopt heterogeneous representations for the same concepts, for example, using numerical labels (e.g., 0 and 1) in some datasets and textual labels (e.g., *hate* and *non-hate*) in others. To reconcile these differences, we applied a series of standardization, and normalization procedures after collecting data from three distinct sources. Labels originally expressed with different formats were mapped to a shared schema consisting of a binary *harmful / not harmful* label, indicating whether a text contains harmful content, and a multi-class *harm subcategory* label with possible values *hate speech*, *stereotype*, *hate speech + stereotype*, and *neutral*. This design allows us to harmonize data across sources while preserving a sufficient level of granularity to support fine-grained analyses of harmful content.

In addition, we assign a *confidence level* to each label to reflect how it was obtained, distinguishing between labels derived from manual annotation, weak supervision, or automated generation. This metadata enabled us a more nuanced analyses of model performance and data quality, and allows researchers to account for varying degrees of label reliability when training or evaluating models.

4.3. *Distribution of the data based on different features*

The resulting framework consists of 237,956 instances described by 18 features, whose categorization and intended purpose for each column are detailed Table 1. In this section, we analyze the distribution of the main features within the dataset and evaluate key data properties, including completeness and linguistic variety.

Table 2 presents the distribution of the two primary annotation dimensions: (i) the binary harmfulness label, which indicates whether a text contains harmful content, and (ii) the harmful language subcategory, which specifies the type of harmful content (e.g., hate speech, stereotype, etc.). The binary distribution shows an approximate 72% / 27% split between harmful and not-harmful instances. This imbalance is expected, as the framework intentionally covers multiple forms of harmful language, requiring a larger number of harmful examples to adequately capture their diversity. Regarding subcategories, nearly half of the harmful instances are labeled as hate speech, which can be attributed to the greater availability of publicly accessible resources and annotated datasets for this category compared to other forms of harmful language.

Table 2. Harmful language and its subcategories distribution in our universal data framework.

harmful_or_not	**Distribution (%)**	**harm_subcategory**	**Distribution (%)**
harmful	72.14	hate_speech	50.47
		hatespeech_stereotype	10.54
		stereotype	9.29
		non_inclusive	0.98
		irony	0.47
		irony_sarcasm	0.38
not_harmful	27.86	neutral	25.27
		inclusive	2.59

Table 3 illustrates the *subcategories* we have in the dataset, which represents the social attribute of the group under attack. Within our framework, approximately 45% of the instances belong to the **gender** category. The predominance of this category can be explained by two main factors. First, the existing literature shows a strong focus on gender-based abuse, with many studies addressing **misogyny** as a primary case study (Frenda et al., 2019). Second, the synthetically generated portion of our dataset related to inclusive language places particular emphasis on gender, due to the inherently gendered nature of the Italian language.

For each category, we also report a subset of the most frequently occurring target groups, which helps clarify how targets are represented in the data and illustrates the meaning of the intersectionality feature. Overall, 86.93% of the data is non-intersectional, meaning that the text targets a single group, while 13.07% is

intersectional, where multiple social groups are targeted simultaneously.

Table 3. Distribution of Target Groups by Category and Intersectionality

Category	Distribution (%)	Target Group(s)	Share (%)	Intersectionality
Gender	52.44	Women only	35.71	Not intersectional
		Women + Immigrants	4.84	Intersectional
		Men only	0.84	Not intersectional
		Women + LGBTQ+	0.65	Intersectional
		Women + Muslims	0.32	Intersectional
Religion	29.68	Jews only	23.48	Not intersectional
		Muslims + Jews	5.47	Intersectional
		Muslims + Immigrants	2.82	Intersectional
		Muslims only	2.66	Not intersectional
		Jews + Immigrants	1.15	Intersectional
		Muslims + Jews + Immigrants	0.73	Intersectional
		Muslims + Women	0.38	Intersectional
		Muslims + Immigrants + Jews*	0.38	Intersectional
Ethnicity	11.47	Immigrants only	9.77	Not intersectional
		Immigrants + Women	0.36	Intersectional
		Immigrants + Muslims	0.32	Intersectional
Disability	4.11	People with disabilities	4.3	Not intersectional
Sexual Orientation	2.3	LGBTQ+ only	2.41	Not intersectional
		LGBTQ+ + Women	0.23	Intersectional

**Note: This appears duplicate but represent different combinations in original data.*

4.4. *Text Complexity and Linguistic Variety*

As described in Section 3.3, three metrics were employed to quantify textual complexity within the proposed framework: the Gulpease Index, lexical diversity, and the sophistication score. Figures 2 to 5 present the average of these metrics across different dimensions of the framework, namely the original data source, harmful language subcategory, category, and the top ten most frequent target groups. Together, these analyses provide insight into how linguistic complexity varies across data characteristics and annotation dimensions.

4.4.1. *Text complexity by data source*

Figure 2 highlights the differences in text complexity across data sources. Curated datasets such as HaSpeeDe and IronITA exhibit higher Gulpease scores and lexical diversity, indicating more readable and lexically varied content, likely due to editorial filtering and annotation guidelines. In contrast, the synthetically generated data exhibit lower readability and reduced linguistic sophistication, reflecting a more uniform and constrained language structure. This effect is attributable to the masking-and-replacement strategy adopted during data generation, which limits lexical variability and promotes repetitive syntactic patterns by substituting specific spans of text with predefined alternatives.

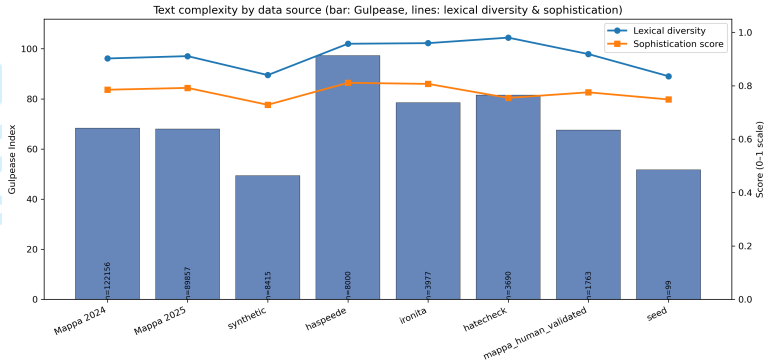

Fig. 2. Text complexity distribution by data source.

4.4.2. *Text complexity by harmful language subcategory*

Figure 3 shows that ironic and irony–sarcasm texts are characterized by higher readability and lexical diversity compared to direct hate speech or stereotype-related content. This suggests that indirect or figurative harmful expressions tend to employ richer and more varied language. Conversely, inclusive and non-inclusive categories display lower Gulpease scores, indicating simpler and more explicit constructions, which is consistent with the synthetic-generated nature of such texts.

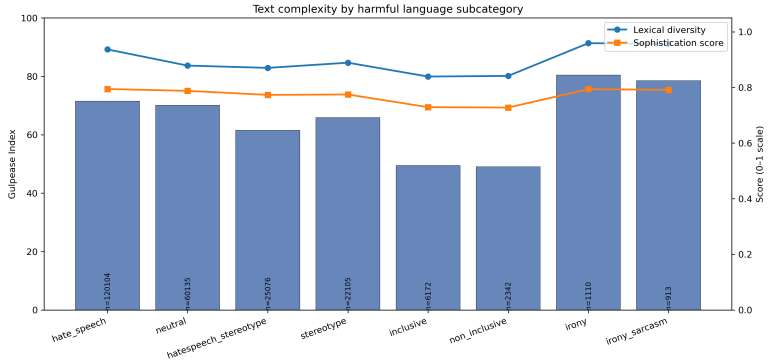

Fig. 3. Text complexity distribution across harmful language subcategories.

4.4.3. *Text complexity by target category*

As shown in Figure 4, texts targeting gender, ethnicity, and religion exhibit relatively similar levels of lexical diversity and sophistication. Overall, the relatively stable trends across categories suggest that text complexity is influenced more by

discourse style than by the target category alone.

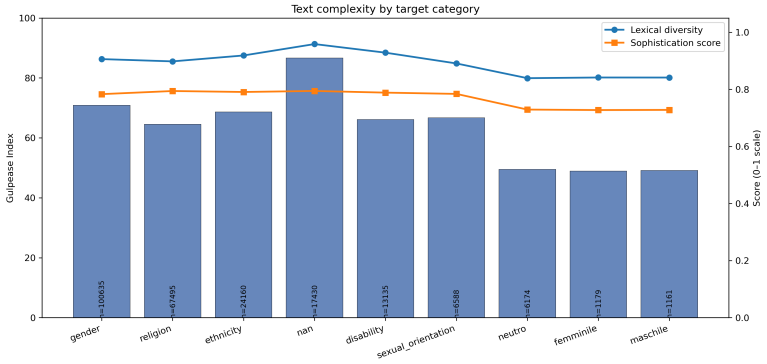

Fig. 4. Text complexity distribution by target category.

4.4.4. *Text complexity across target groups*

Figure 5 provides a more fine-grained view by examining the ten most frequent target groups, including single and intersectional targets. Mixed targets such as "jews + muslims" or "immigrants + women" tend to exhibit lower Gulpease scores and reduced lexical diversity, indicating more formulaic and repetitive language patterns. In contrast, single-target groups such as "women" or "disabled" show higher complexity values. This observation suggests that intersectional harmful content may rely on more stereotyped and less linguistically diverse expressions.

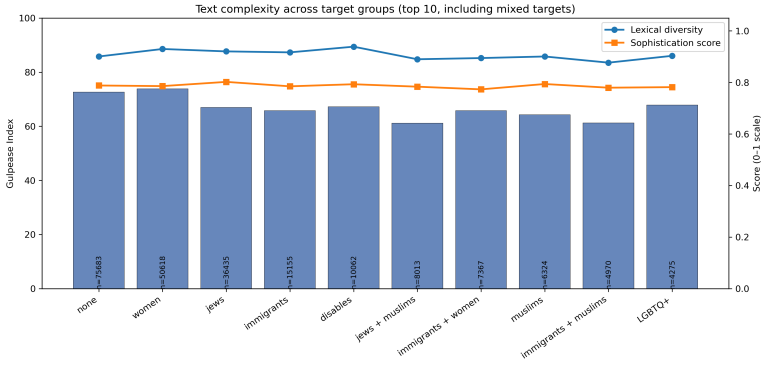

Fig. 5. Text complexity distribution across target groups.

4.4.5. *Linguistic Variety Assessment*

Another metric employed to assess the quality of the data collected in the framework is linguistic variety, measured using the Self-BLEU metric (Zhu et al., 2018). To compute this metric, we randomly sampled 1% of the overall dataset and calculated Self-BLEU scores for three different n-gram sizes, namely n=1, 2, and 4. The results indicate a high level of linguistic diversity within the dataset. Specifically, the Self-BLEU score decreases from 0.7591 at the unigram level, which reflects substantial lexical overlap across sentences, to 0.1073 at the 4-gram level, where similarity is measured over longer phrase sequences. This sharp decrease demonstrates that, while individual words may recur frequently, longer syntactic and semantic patterns are considerably more varied, suggesting a low level of redundancy and a high degree of textual diversity in the framework.

4.5. *Model Selection for Downstream Tasks*

Once the data have been collected, cleaned through quality checks, normalized, integrated into the framework, and enriched with additional features, they are used for model inference. This stage of the workflow aims to evaluate model performance on the target downstream tasks and to iteratively improve results through controlled adaptations of both data and training strategies.

4.5.1. *Baseline Inference*

To establish a reference point for evaluation, we first perform baseline inference using publicly available open-source large language models obtained from repositories such as Hugging Face. These models are evaluated without any task-specific adaptation in order to measure their out-of-the-box performance. The resulting metrics serve as a benchmark against which models are compared and allow us to identify the most promising architectures for subsequent adaptation. Table 4 illustrates the results we achieved with our dataset.

Table 4. Baseline student models on binary harmful language detection

Model	**Acc**	**Macro**			**Weighted**		
		P	**R**	**F1**	**P**	**R**	**F1**
Gemma-3-4b-it	0.5831	0.7608	0.5637	0.4729	0.7517	0.5831	0.4837
Phi-4	0.6693	0.7799	0.6546	0.6209	0.7725	0.6693	0.6270
Ministral-3B	0.6168	0.7340	0.5910	0.5324	0.7240	0.6168	0.5457
Qwen3-4B	0.7157	0.7176	0.7125	0.7127	0.7171	0.7157	0.7140

4.5.2. *Dataset Formatting and Prompting*

Prompts are defined using a consistent template that provides task context, assigns a clear role to the model, and specifies the classification objective. Initial experiments are conducted using zero-shot prompts, while more advanced strategies, such as few-shot prompting and chain-of-thought–inspired formulations, are explored in later stages to improve output consistency and task understanding. Selected datasets are converted from their original tabular format into an instruction-following conversational structure compatible with the target models.

4.5.3. *Iterative Design and Evaluation Cycle*

Model selection and adaptation follows an iterative design lifecycle driven by integrated data-centric and model-centric evaluations. At each iteration, we systematically conduct:

- **distributional analysis** to identify class imbalance, source bias, and representation gaps,
- **linguistic diversity assessment** to ensure adequate coverage across lexical, syntactic, and stylistic variations,
- **performance evaluation** on the target downstream task using established metrics.

When performance outcomes doesn't meet the expectations, we do not rely solely on model slection and configuration. Instead, we implement targeted, data-level interventions such as rebalancing class distributions, adjusting the contribution weights of different data sources by changing the *label_confidence* values, and enriching linguistic diversity in underrepresented areas. This closed-loop process enables us to improve performance systematically while maintaining full transparency and rigorous control over data quality throughout the development cycle.

4.6. *Evaluation*

Model evaluation is conducted after completing the iterative design and evaluation cycle described in the previous section. Through successive rounds of data re balancing, linguistic enrichment, and prompt design, GPT-5-nano emerged as the most effective model in terms of both predictive performance and computational efficiency. As such, GPT-5-nano is selected as the reference model for the final evaluation and comparison with existing approaches.

Table 5 reports the per-class performance of GPT-5-nano on the binary harmful language detection task. The model achieves a high recall of 0.9283 for the `harmful` class, resulting in an F1-score of 0.8133. This indicates a strong capability to identify harmful content, which is a critical requirement in moderation-oriented scenarios. Conversely, the `not_harmful` class exhibits higher precision (0.8862) but lower recall

(0.6119), reflecting a conservative bias toward minimizing false negatives for harmful content.

The selection of GPT-5-nano is the outcome of the iterative optimization process, in which performance improvements were achieved not only through model-centric tuning but also through targeted data-centric interventions. In particular, adjustments to data distributions, refinement of *label_confidence* weights across heterogeneous sources, and increased linguistic diversity in underrepresented patterns contributed to more stable and consistent performance across evaluation runs.

To contextualize these results, we compare GPT-5-nano with prior work on Italian harmful and hate speech detection, as summarized in Table 5. In the MULTA-TELEGRAM benchmark (Leonardelli et al., 2025), in-domain BERT-based multi-task models report an F1-score of 0.732 ± 0.025, while LLaMA-based models achieve F1-scores of 0.732. General-purpose safety models such as LLaMA-Guard achieve an F1-score of 0.712, highlighting the limitations of broad safety-oriented systems when applied to task-specific detection.

More recent large-scale models achieve higher absolute scores but at significantly higher computational cost. For example, (Magnini et al., 2025) report an F1-score of 0.77 for Qwen2.5-14B-Instruct-1M on hate speech detection, while top-performing systems in the HaSpeeDe 2 challenge at EVALITA (Del Vigna12 et al., 2017) achieve F1-macro scores of 0.80 on Twitter data and 0.77 on news headlines. These results are obtained in competition settings using substantially larger models and task-specific training regimes.

Overall, the GPT-5-nano that we selected using our framework and adapted through prompt engineering outperforms the available baselines in the literature for the Italian language. This demonstrates the effectiveness of our proposed iterative, data-centric design methodology.

5. Conclusion

This paper introduces **EHLD**, a general and extensible data framework designed to reduce fragmentation in harmful language detection research. It enables the integration of heterogeneous datasets into a unified, analysis-ready schema. EHLD combines essential detection labels with optional contextual attributes and **evaluation-oriented features** that capture text complexity. It also preserves provenance and supports reproducibility through its metadata. We instantiated the framework by integrating curated datasets from the literature, large-scale LLM-based annotation from the *Mappa dell'Intolleranza*, and LLM-labeled data for inclusive language detection, resulting in a consolidated dataset of 237,956 Italian instances described by 18 features.

A key outcome of the proposed methodology is the adoption of an **iterative, data-centric design** and evaluation cycle. In this cycle, model performance is improved through prompt design and model selection, as well as through systematic interventions on the data itself. Specifically, we analysed distributional properties

Table 5. Comparison of GPT-5-nano with prior work on Italian harmful/hate speech detection.

Model	Task	F1	Notes
Our approach			
			Selected using the EHLD framework
GPT-5-nano	Harmful language (binary)	**0.813**	Selected after iterative data-centric optimization
Related work			
			MuLTa-Telegram (Leonardelli et al., 2025)
BERT-based	Harmful language	0.732 ± 0.025	Manually annotated
LLaMA	Harmful language	0.732	General LLM baseline
LLaMA-Guard	Harmful language	0.712	General-purpose safety-focused model
			Benchmarking LLMs on Italian (Magnini et al., 2025)
Qwen2.5-14B	Hate speech	0.77	Instruction-tuned LLM
			HaSpeeDe 2 (Del Vigna12 et al., 2017)
BERT-based	Hate speech (Twitter)	0.80	Competition setting
BERT-based	Hate speech (Headlines)	0.77	Competition setting

to detect imbalances and gaps in representation, assessed **linguistic diversity** to identify underrepresented patterns, and revised the composition of the dataset by rebalancing and adjusting the contributions of the sources through *label_confidence*. Within this controlled loop, GPT-5-nano emerged as the most effective model under our constraints, offering an excellent balance of accuracy and efficiency.

In terms of performance, **GPT-5-nano** achieves an F1 score of 0.813 for binary *harmful* language detection, with a particularly high recall of 0.928 for the harmful class. While this behaviour is desirable in moderation-oriented settings where missing harmful content is risky, it is accompanied by lower recall for the *not_harmful* class, suggesting a conservative decision boundary. When contextualised against the results reported in the literature, the performance obtained exceeds the **baseline models** reported for Italian harmful language detection in comparable settings, such as BERT-based multitask approaches and other LLMs. Importantly, EHLD facilitates more interpretable comparisons by explicitly tracking provenance and label reliability, and by providing evaluation-oriented features that help explain when and why models succeed or fail.

Despite these contributions, several **limitations** remain. First, performance comparisons across studies are inherently constrained by differences in datasets,

domains (e.g., tweets vs. headlines), annotation guidelines, and reported metrics. Second, although EHLD encodes label reliability via *label_confidence*, a substantial portion of the data originates from automatic annotation pipelines, which may contain residual noise that impacts both training and evaluation. Thirdly, while the schema is language-agnostic, the current instantiation is centred on Italian, and **cross-lingual validation** and multilingual integration have not yet been empirically demonstrated within this paper. Finally, the presented evaluation focuses on binary harmfulness, leaving robust multi-class and fine-grained target-level prediction as open challenges.

Future research can extend EHLD along two complementary axes. From a data perspective, richer coverage of underrepresented harmful phenomena (e.g., implicit abuse, coded language, and intersectional targeting) can be achieved through targeted data collection, expert validation, and the systematic enrichment of missing contextual attributes. From an evaluation perspective, EHLD can be expanded to support **fine-grained evaluation metrics** that go beyond aggregate scores, including detailed per-class precision, recall, and F1 scores. In addition, incorporating advanced deployment-oriented metrics such as the **abstention level** can better characterise model reliability under uncertainty and support risk-aware assessment in safety-critical moderation settings. Overall, EHLD provides a foundation for more interoperable datasets and transparent evaluation practices, supporting cumulative progress towards robust systems for detecting harmful language.

Acknowledgment

The work reported in this paper has been partly funded by the European Union - NextGenerationEU, under the National Recovery and Resilience Plan (NRRP) Mission 4 Component 2 Investment Line 1.5: Strengthening of research structures and creation of R&D "innovation ecosystems", set up of "territorial leaders in R&D", within the project "MUSA - Multilayered Urban Sustainability Action" (contract no. ECS 00000037).

References

Alaçam, Ö., Hoeken, S., Säuberli, A., Gröner, H., Frassinelli, D., Zarrieß, S., and Plank, B. (2025). Disentangling subjectivity and uncertainty for hate speech annotation and modeling using gaze. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28695–28712.

Albladi, A., Islam, M., Das, A., Bigonah, M., Zhang, Z., Jamshidi, F., Rahgouy, M., Raychawdhary, N., Marghitu, D., and Seals, C. (2025). Hate speech detection using large language models: A comprehensive review. *IEEE Access*.

Antypas, D. and Camacho-Collados, J. (2023). Robust hate speech detection in social media: A cross-dataset empirical evaluation. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Arango, A., Pérez, J., and Poblete, B. (2019). Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, pages 45–54.

Arora, A., Nakov, P., Hardalov, M., Sarwar, S. M., Nayak, V., Dinkov, Y., Zlatkova, D., Dent, K., Bhatawdekar, A., Bouchard, G., et al. (2023). Detecting harmful content on online platforms: what platforms need vs. where research efforts go. *ACM Computing Surveys*, 56(3):1–17.

Artstein, R. (2017). Inter-annotator agreement. In *Handbook of linguistic annotation*, pages 297–313. Springer.

Baiocco, R., Rosati, F., and Pistella, J. (2023). Italian proposal for non-binary and inclusive language: The schwa as a non-gender–specific ending. *Journal of Gay & Lesbian Mental Health*, 27(3):248–253.

Bellini, E. (2025). Cyber humanities for heritage security. *Commun. ACM*, 68(12):112–117.

Bosco, C., Dell'Orletta, F., Poletto, F., Sanguinetti, M., Tesconi, M., et al. (2018). Overview of the evalita 2018 hate speech detection task. In *Ceur workshop proceedings*, volume 2263, pages 1–9. CEUR.

D'Amico, M. et al. (2023). *Parole che separano: linguaggio, Costituzione, diritti*, volume 151. Raffaello Cortina.

Del Vigna12, F., Cimino23, A., Dell'Orletta, F., Petrocchi, M., and Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the first Italian conference on cybersecurity (ITASEC17)*, pages 86–95.

D'Amico, M. and Siccardi, C. (2020). *VERSO UNA CULTURA CHE NON ODIA*. Giapichelli Editore, Milan, Italy.

Fiano, N. (2025). The "mappa dell'intolleranza" project. In *Book of Abstracts*, page 276.

Fortuna, P., Soler, J., and Wanner, L. (2020). Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6786–6794.

Frenda, S. (2023). Creative and cognitive aspects of implicitness in abusive language online. *Natural Language Engineering*, 29(6):1516–1537.

Frenda, S., Ghanem, B., Montes-y Gómez, M., and Rosso, P. (2019). Online hate speech against women: Automatic identification of misogyny and sexism on twitter. *Journal of intelligent & fuzzy systems*, 36(5):4743–4752.

Gong, Y., Liu, G., Xue, Y., Li, R., and Meng, L. (2023). A survey on dataset quality in machine learning. *Information and Software Technology*, 162:107268.

Kiritchenko, S., Nejadgholi, I., and Fraser, K. C. (2021). Confronting abusive language online: A survey from the ethical and human rights perspective. *Journal of Artificial Intelligence Research*, 71:431–478.

Korre, K., Pavlopoulos, J., Sorensen, J., Laugier, L., Androutsopoulos, I., Dixon,

L., and Barrón-Cedeño, A. (2023). Harmful language datasets: An assessment of robustness. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 221–230.

Kumar, M. et al. (2025). Exploring hate speech detection: challenges, resources, current research and future directions. *Multimedia Tools and Applications*, pages 1–37.

Leonardelli, E., Casula, C., Salto, S. V., Bak, J. E., Muratore, E., Kołos, A., Louf, T., and Tonelli, S. (2025). Multa-telegram: A fine-grained italian and polish dataset for hate speech and target detection. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 582–594.

Lucisano, P. and Piemontese, M. (1988). Gulpease: una formula per la predizione della difficoltà dei testi in lingua italiana [gulpease: a formula to predict the difficulty of texts in italian language], in scuola e città, 3. *La Nuova Italia*.

Maghool, S., Mohammadi, F., Ceravolo, P., and Damiani, E. (2025). From seed to synthetic: A data generation protocol for fair, inclusive, and robust language detection. Under review at IEEE Access.

Magnini, B., Cappelli, A., Pianta, E., Speranza, M., Bartalesi Lenzi, V., Sprugnoli, R., Romano, L., Girardi, C., and Negri, M. (2006). Annotazione di contenuti concettuali in un corpus italiano: I - cab. In *Proc.of SILFI 2006*.

Magnini, B., Madeddu, M., Resta, M., Zanoli, R., Cimmino, M., Albano, P., and Patti, V. (2025). A leaderboard for benchmarking llms on italian. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 636–646.

Mahomed, Y., Crawford, C. M., Gautam, S., Friedler, S. A., and Metaxa, D. (2024). Auditing gpt's content moderation guardrails: Can chatgpt write your favorite tv show? In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 660–686.

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., and Patti, V. (2021). Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55(2):477–523.

Polignano, M., Basile, P., De Gemmis, M., Semeraro, G., and Basile, V. (2019). Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*, pages 312–317.

Rauh, M., Mellor, J., Uesato, J., Huang, P.-S., Welbl, J., Weidinger, L., Dathathri, S., Glaese, A., Irving, G., Gabriel, I., et al. (2022). Characteristics of harmful text: Towards rigorous benchmarking of language models. *Advances in Neural Information Processing Systems*, 35:24720–24739.

Romano, T., Mohammadi, F., and Ceravolo, P. (2024). Synthetic data for identifying inclusive language (case study: Job descriptions in italian). In *2024 IEEE International Conference on Cyber Security and Resilience (CSR)*, pages 737–742. IEEE.

Röttger, P., Vidgen, B., Nguyen, D., Talat, Z., Margetts, H., and Pierrehumbert, J. (2021). Hatecheck: Functional tests for hate speech detection models. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pages 41–58.

Tamborini, M. A., Mohammadi, F., Ceravolo, P., Maghool, S., D'Amico, M. E., and Siccardi, C. (2025). Agent for anti-discriminatory language (aal) a human-centered ai approach to discriminatory speech in italian. In *2025 IEEE International Conference on Cyber Humanities (IEEE-CH)*, pages 1–6.

Zhang, C., Zhu, C., Xiong, J., Xu, X., Li, L., Liu, Y., and Lu, Z. (2025). Guardians and offenders: A survey on harmful content generation and safety mitigation of llm. *arXiv preprint arXiv:2508.05775*.

Zhou, Z.-H. (2018). A brief introduction to weakly supervised learning. *National science review*, 5(1):44–53.

Zhu, Y., Lu, S., Zheng, L., Guo, J., Zhang, W., Wang, J., and Yu, Y. (2018). Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1097–1100.

Biography