POLITECNICO DI TORINO
Repository ISTITUZIONALE

Towards AI-Assisted Inclusive Language Writing in Italian Formal Communications

Original
Towards AI-Assisted Inclusive Language Writing in Italian Formal Communications / Greco, S., La Quatra, M., Cagliero, L., Cerquitelli, T.. - In: ACM TRANSACTIONS ON INTELLIGENT SYSTEMS AND TECHNOLOGY. - ISSN 2157-6904. - ELETTRONICO. - 16:4(2025), pp. 1-24. [10.1145/3729237]

Availability:
This version is available at: 11583/3000852 since: 2025-06-11T09:16:40Z

Publisher:
Association for Computing Machinery

Published
DOI:10.1145/3729237

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

23 November 2026



Towards AI-Assisted Inclusive Language Writing in Italian Formal Communications

SALVATORE GRECO, Politecnico di Torino, Turin, Italy
MORENO LA QUATRA, Kore University of Enna, Enna, Italy
LUCA CAGLIERO and TANIA CERQUITELLI, Politecnico di Torino, Turin, Italy

Formal communications such as public calls, announcements, or regulations are supposed to exhibit respect for diversity in terms of gender, race, age, and disability. However, human writers often lack adequate inclusive writing skills. For instance, they tend to overuse the masculine as a neutral form, mainly because they are self-trained on biased text examples. To overcome this issue, we propose to leverage Generative Artificial Intelligence to support inclusive language writing. Focusing on formal Italian communications, we have designed and developed an AI-assisted tool for non-inclusive text detection and reformulation. Thanks to the joint work with a team of linguistic experts, we first define a set of linguistic criteria necessary to model inclusive writing forms in Italian. Based on these criteria, we collect and annotate a dataset of Italian administrative documents enriched with fine-grained inclusive annotations. Finally, we train deep learning models on the collected data for non-inclusive language detection and inclusive language reformulation tasks. We perform quantitative and human-driven evaluations on the trained models. The best detection model correctly classifies 89% of the sentences, whereas the best reformulation model produces 73% fully correct reformulations. Both models have been integrated into a writing assistance tool acting as a text proofreader and self-learning tool for non-expert writers, namely Inclusively. Once a non-inclusive piece of text is detected, the proposed approach suggests inclusive reformulations. The tool also provides explanations of the models' outputs to increase system transparency. Furthermore, it allows expert end-users to provide further annotations for system fine-tuning. The trained models and the writing assistance tool are publicly available for research purposes.

CCS Concepts: • **Computing methodologies** → **Natural language generation**; **Natural language processing**;

Additional Key Words and Phrases: inclusive language, natural language processing, text classification, text generation

Salvatore Greco and Moreno La Quatra contributed equally to this research.

This study was carried out within the project "E-MIMIC: Empowering Multilingual Inclusive Communication" (Grant No. 2022WEFCFP), funded by the Ministero dell'Universitá e della Ricerca—with the PRIN 2022 (D.D. 104–02/02/2022) program. This article reflects only the authors' views and opinions, and the Ministry cannot be considered responsible for them.

Authors' Contact Information: Salvatore Greco (corresponding author), Politecnico di Torino, Torino, Italy; e-mail: salvatore_greco@polito.it; Moreno La Quatra, Kore University of Enna, Enna, Italy; e-mail: moreno.laquatra@unikore.it; Luca Cagliero, Politecnico di Torino, Torino, Italy; e-mail: luca.cagliero@polito.it; Tania Cerquitelli, Politecnico di Torino, Torino, Italy; e-mail: tania.cerquitelli@polito.it.



This work is licensed under Creative Commons Attribution International 4.0.

© 2025 Copyright held by the owner/author(s).

ACM 2157-6912/2025/6-ART79
https://doi.org/10.1145/3729237

ACM Reference format:
Salvatore Greco, Moreno La Quatra, Luca Cagliero, and Tania Cerquitelli. 2025. Towards AI-Assisted Inclusive Language Writing in Italian Formal Communications. *ACM Trans. Intell. Syst. Technol.* 16, 4, Article 79 (June 2025), 24 pages.
https://doi.org/10.1145/3729237

1 Introduction

Effective use of language is fundamental for sharing thoughts, declaring laws or regulations, and communicating news feeds. Inclusivity represents a crucial linguistic feature in achieving effective communication. The use of inclusive language holds particular relevance in formal communications, such as public calls, minutes, and administrative messages, to ensure that all individuals feel acknowledged, respected, and treated equally regardless of gender, race, age, disability, and belief [49].

As a matter of fact, most people, including administrative office employees, lack the necessary skills in inclusive writing. The motivations behind this are manifold [6]. Firstly, they underestimate the negative impact of non-inclusive writing, which might convey biased, exclusive, or offensive messages. Secondly, organizations often fail to offer sufficient support or training, letting employees rely on self-learning. Lastly, the prevailing examples of formal communication contain non-inclusive elements, thereby perpetuating bad writing habits [22].

The goal of this work is to support the activity of inclusive writing by proposing a tool that can help eliminate stereotypes of gender, ethnicity, disability, and age present in an input document. Specifically, we focus on inclusive language in Italian formal communications. We choose the formal communication domain due to its inherent high level of formality, which necessitates primarily an inclusive writing style. Additionally, the Italian language, like other Romance languages, is characterized by an extensive presence of gendered language [33].

An example of a gendered expression in Italian is "*The students*," which can be written in Italian either in the masculine form as "*Gli studenti*" or in the feminine form as "*Le studentesse*." However, the common writing practice is to use the masculine form as a neutral option, which does not promote inclusivity. A more inclusive reformulation would be to adopt a gender-neutral term such as "*La componente studentesca*" (EN: "*The student body*"). Unfortunately, gendered expressions like the one mentioned above are likely to occur in formal communications.

To support adopting inclusive forms, we envisage the application scenario depicted in the bottom pipeline of Figure 1. It incorporates a pipeline consisting of two main components: (1) A *classifier*, which acts as a non-inclusive text detector and (2) A *generative model*, which serves as the text reformulation model. Let us consider a non-linguistics expert responsible for composing formal documents. To meet inclusive language standards, the initial draft of the document can be automatically reviewed and revised by an AI-based writing assistant tool, namely INCLUSIVELY. The tool provides end-users with a writing assistance interface that first scans the document to identify snippets of text lacking inclusivity. Next, it leverages generative modeling to reformulate these non-inclusive portions in a gender-neutral manner while prioritizing accessibility. This approach guarantees that the text is easily understood by individuals with disabilities or visual impairments without resorting to special letters or symbols that might hinder accessibility.

When the suggested reformulation is unsatisfactory, the tool provides an interface that allows expert users to manually provide additional annotations for further system refinement (i.e., evaluation and annotation interface). Finally, to facilitate both the evaluation of the generated outputs and the self-training of human writers, the tool provides end-users with explanations of the models'

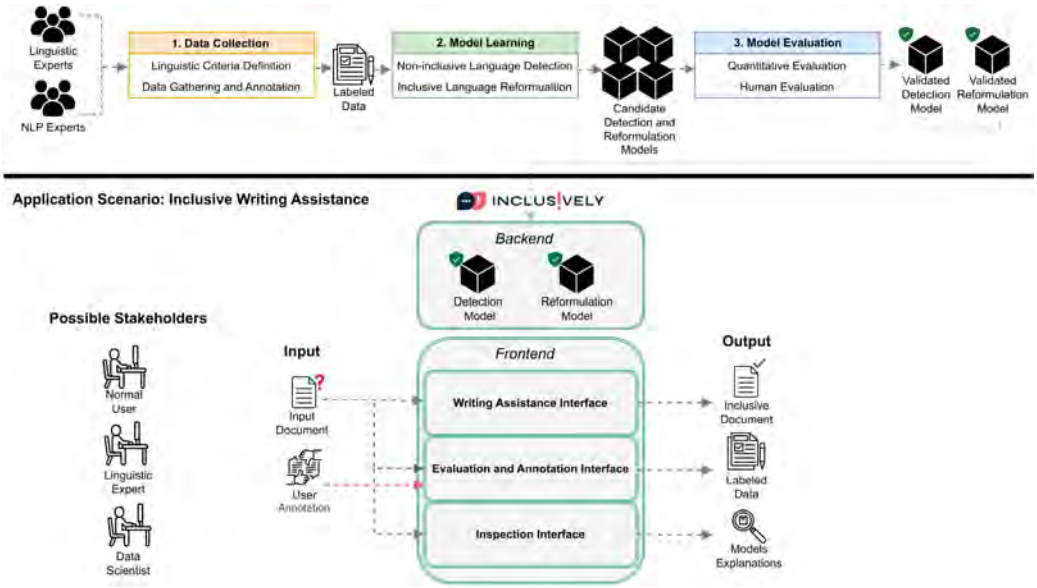


Fig. 1. *Methodology and application scenario overview.* The data-driven methodology to model inclusive language consists of three steps. (1) *Data collection:* Linguistic experts defined the linguistic criteria to model inclusive language, and gathered and annotated a corpus of Italian formal documents to produce a labeled dataset; (2) *Model learning*: We trained transformer-based classifiers and sequence-to-sequence models to detect and rephrase non-inclusive language; and (3) *Model evaluation*: We perform quantitative and human validation of the trained models. We employed the best-performing models in the writing assistance tool, namely INCLUSIVELY. The tool presents three interfaces: (i) *Writing assistance interface* where normal users can input a formal document, and the tool identifies and reformulates segments of text that lack inclusivity; (ii) *Evaluation and annotation interface* where linguistic experts can manually provide additional annotations and feedback for further system refinement; and (iii) *Inspection interface* which provides end-users with explanations of the models' outputs.

outputs (through the inspection interface). End-users can gain relevant insights into the AI model reasoning, which is typically not transparent [9].

The proposed solution offers several advantages: (1) It relieves linguistic experts from having to manually review the raw content, thus making inclusive writing less expensive and more affordable, even in the public domain; (2) It leverages the increasing power of Generative AI tools such as GPT-based Language Models [7, 36] to generate inclusive versions of inappropriate expressions; and (3) It empowers non-expert writers to self-train on the correct writing style and, hopefully, prevent future inaccuracies in writing.

The primary issue in developing the tool revolves around the lack of annotated data and models specifically tailored for inclusive language tasks in the Italian context. To address this challenge, this article presents the data-driven methodology, discusses the empirical findings, and outlines the evaluation process we conducted to develop the powerful detection and reformulation models utilized within the tool [24], as depicted in the upper pipeline of Figure 1. In doing so, our contributions are threefold:

(1) *Guidelines Definition.* Thanks to joint work with linguistic experts, we first devised a set of linguistic criteria for the Italian inclusive language. Based on these criteria, we then collect

and annotate a real corpus of Italian administrative texts for non-inclusive language detection and reformulation tasks (Section 3).

(2) *Human Validation Design.* We design two human evaluation tasks to assess the outputs of the detection and reformulation models (Section 4). To evaluate the detection models, we design a task where linguistic experts assess if the models use the correct words for the predictions of the inclusive label by exploiting explainability techniques. To evaluate the reformulation models, linguistic experts check whether the proposed reformulations solve all the inclusivity issues, maintain grammatical correctness, and preserve the original text's meaning.

(3) *Model Learning and Evaluation.* We fine-tune two transformer-based text classifiers to detect non-inclusive sentences and two sequence-to-sequence models to rephrase the non-inclusive forms into inclusive ones. Then, we perform an extensive evaluation of the quality of their outcomes using both standard quantitative metrics and the designed human-driven tasks (Section 5). Quantitatively, the best performing detection model correctly discriminates between inclusive and non-inclusive sentences in 9 out of 10 cases. Linguistic experts verify that this model correctly uses only the words determining the inclusiveness or not of the sentence in 87% of the cases. The best reformulation model achieves high-quality quantitative performance according to the reference metrics (i.e., **Bilingual Evaluation Understudy (BLEU)** [30]: 0.81, **Recall-Oriented Understudy for Gisting Evaluation (ROUGE)**-2 F1 score [26]: 0.87). According to the experts' judgment, enriching the collection with template-based examples significantly enhances the quality of the text reformulations. Overall, the best reformulation model produces correct reformulations (i.e., inclusive reformulations that preserve the original meaning and maintain grammatical correctness) in 73% of the cases.

To foster future research and inclusive language writing, we release the tool source code and the trained models.[1]

2 Background and Related Work

In this section, we first review the available **Natural Language Processing (NLP)** resources for the Italian language (Section 2.1). We then introduce the problem of inclusive language and the guidelines published by different institutions worldwide, with a particular focus on the Italian language (Section 2.2). Finally, we review NLP studies on inclusive intra-language understanding and generation (Section 2.3).

2.1 Italian Language Modeling

Developing deep learning-based approaches to understanding Italian documents is particularly challenging due to the inherent complexity of the language and the limited number of available resources (e.g., datasets, vocabularies, knowledge bases, pre-trained models) compared to more widely spoken languages such as English. The Italian NLP community has recently started to bridge this gap. For example, in [34, 45], the authors propose pre-trained BERT models [14] for general-purpose and social media language understanding tasks, respectively. They leverage self-supervised pre-training objectives to learn high-level representations from large amounts of unlabeled data. These representations can be fine-tuned to perform various NLP tasks (e.g., text classification).

Lately, Italian sequence-to-sequence models have also been presented. For instance, IT5 [40] is a pre-trained encoder-decoder model reproducing the original T5 architecture [37]. The model is pre-trained on a cleaned version of the Italian text contained in C4 corpus [37] using multi-task

[1] https://github.com/MorenoLaQuatra/inclusively.

learning. Aiming to learn language-specific representations, IT5 is trained only using Italian-written documents. The authors show that the model can be fine-tuned for tasks such as text summarization and question answering. Similarly, BART-IT [23], a pre-trained version of BART for the Italian language, is trained on the same cleaned corpus of Italian text and has demonstrated robust performance, particularly in text summarization.

Some attempts to develop and test decoder-only models for the Italian language have also been made. For instance, GePpeTto [28] is a GPT-2-like model pre-trained from scratch on an Italian corpus. The model can be used to generate fluent and coherent text.

All these models can be fine-tuned to downstream tasks using labeled data. In our specific case, the objective is to fine-tune such models for non-inclusive language detection and reformulation. However, to the best of our knowledge, an Italian dataset labeled for addressing these particular tasks is not available yet.

2.2 Inclusive Language Guidelines

In recent years, many institutions have emphasized the importance of adopting inclusive language. For instance, in the U.S., the American Psychological Association has established a set of guidelines for inclusive writing. These guidelines aim to eliminate discrimination and other barriers contributing to unequal opportunities [3]. The United Nations defined a set of principles for adopting an inclusive approach to everyday writing practices [29]. In the UK, the government and other companies delivered some practices for the use of language for gender-neutral writing [20] and disability [19, 51]. Finally, the European Parliament released a set of guidelines used in their formal communications [31]. Other countries and institutions have released their (language-dependent) guidelines and rules. These guidelines raise the importance of using inclusive language, especially in formal communication. In this article, we focus on inclusive language in Italian formal communications.

In Italy, numerous public organizations and universities have published their inclusive language guidelines [10–13, 15–17, 32]. The problem is particularly challenging for the Italian language because, like other romance languages, it is highly gendered [33]. As we will discuss in Section 3, in this work, we engage a team of linguistic experts who devised a set of criteria based on their expertise. These criteria are employed throughout the annotation process. In delineating the linguistic criteria, our objective was to promote inclusivity across gender, age, race, and disability aspects rather than exclusively focusing on gender. For instance, we avoid inclusive forms which involve the double feminine masculine form separated by slash (/), e.g., "*Colleghi/e*" and "*Colleghi/Colleghe*" (EN: "*Male colleagues/female colleagues*") or the use of the schwa (ə), e.g., "*Collegha*," in contrast with suggested by some guidelines [16]. Such reformulations may pose challenges for automatic reading tools, rendering them less inclusive for individuals with disabilities.

Despite the presence of numerous guidelines emphasizing the importance of embracing inclusive writing, there is still inadequate adoption of inclusive language in formal communications. One possible reason is that adopting these guidelines can be complex and time-consuming, especially for individuals lacking expertise in linguistics, such as administrative office employees. To foster the use of inclusive language and simplify the writing task, we propose an AI-based writing assistance tool to detect non-inclusive language and propose inclusive reformulations in formal communications.

2.3 Inclusive Language Understanding and Generation

According to their primary objectives, related studies can be classified into two categories: (1) *cross-linguistic*, which address the inclusivity problem for tasks involving two languages such as machine translation (e.g., [33, 41–43]) and (2) *intra-linguistic*, which leverage NLP models to reformulate non-inclusive text into its inclusive version within the same language. Our work falls

Table 1. State-of-the-Art Literature on Inclusive Language Processing

	Task	Supported languages
[1]	NLP-based gender rewriting	Arabic
[2]	Round-trip translation (from gender-neural to gender-biased) and neural text paraphrasing	German
[18]	Rule-based gender rewriting	German
[48]	Rule-based and neural text paraphrasing (from gendered to gender-neutral)	English
[52]	Rule-based and neural text paraphrasing (from gendered to gender-neutral)	English
[54]	Rule-based and neural text paraphrasing (from gendered to gender-neutral)	Portuguese
Inclusively (ours)	Neural non-inclusive language detection and paraphrasing	Italian

into the *intra-linguistic* category. Therefore, in this section, we focus on the intra-linguistic studies, as cross-linguistic approaches are not suitable for assisting writers in a single language, where both the source and output sentences belong to the same language.

Recent surveys have extensively reviewed gender bias and inclusivity challenges across various NLP tasks [46, 47]. Building on this foundation, most relevant efforts have been devoted to tackling gender issues, e.g., by adopting either the *they* neutral form in English [48, 52] or the gender form desired by end-users [1, 18]. To the best of our knowledge, this article is the first attempt to use intra-linguistic approaches to Italian documents. Table 1 summarizes the existing literature on intra-linguistic inclusive language understanding and generation, and the main differences with existing intra-linguistic works are enumerated below.

- —The inherent complexity of the Italian language hinders the use of rule-based approaches that are commonly applied to English documents, e.g., [18, 52].
- —Unlike [2], our approach does not require a corpus of text already written in an inclusive language, which is usually hard to collect since inclusive language is still not widespread.
- —Compared to [48, 52, 54], we not only address gender-related inclusivity issues, but we also support more complex types of reformulations using techniques such as omission, metonymy, and epicene words, which are not easily applicable with rule-based reformulations.

The present work is part of the E-MIMIC project [4, 24, 38].[2] Some preliminary results of non-inclusive language detection using template-based generated data have been presented in [4]. Specifically, in [4], we performed preliminary experiments on a binary non-inclusive language detection task—predicting if a sentence is inclusive or non-inclusive. Due to the lack of annotated datasets, the analysis was conducted on template-generated sentences instead of real-world administrative data. In [4], we show the potential of the approach and the effectiveness of using deep learning-based NLP classifiers in identifying non-inclusive sentences in the Italian language. The current work extends its preliminary version [4] to a large extent. Specifically: (1) It collects and annotates a real-world labeled dataset according to inclusive linguistic criteria defined by linguistic experts instead of relying on template-generated sentences; (2) It addresses and investigates not only the non-inclusive language detection (classification) but also the reformulation tasks by training classification and sequence-to-sequence models on the properly collected data; (3) It designs two human evaluation tasks tailored to non-inclusive language detection and reformulation; and (4) It reports extensive quantitative and human evaluations of the trained models. The best-performing

[2]The project Web site is available at https://dbdmg.polito.it/e-mimic/index.php.

detection and reformulation models based on the empirical evaluation performed in this article are employed in a writing assistant tool. A demo of the tool can be found in [24].

3 Inclusive Dataset: The Italian Case

We have designed and utilized a novel dataset for inclusive writing in Italian. This dataset comprises administrative communications sourced from the Italian public administration, spanning across both national and regional levels. The motivations behind our contribution are enumerated as follows:

— *Why do we focus on the Italian language?* Like other Romance languages, such as Spanish and French, Italian grammar envisages *ad hoc* rules for many categories of inclusiveness, especially for gender. Regarding gender, Italian has masculine and feminine declensions of words (unlike in English). The standard practice so far is to use the masculine form as the neutral form in official communication. However, this practice conflicts with the prevailing guidelines for inclusive language writing. Therefore, leveraging AI to support inclusive Italian writing is particularly appealing [8].
— *Why do we consider official communications?* Whenever a public or private institution issues a formal communication, such as a call, an announcement, or a minute, it is intended for the whole population, including minorities. Therefore, official communications are expected to be as inclusive as possible to embrace and respect diversity.
— *Why do we rely on expert-annotated data?* According to the state-of-the-art literature (see Section 2), there is a lack of annotated data suited to AI-based inclusive language writing, particularly in the Italian language.

3.1 Annotators Team

We set up a team of 13 linguistic experts in the field of inclusive language to collect and annotate the Italian dataset. This team devised specific linguistic criteria for promoting inclusivity within the Italian language. Then, they meticulously collected documents highlighting pertinent inclusivity issues and formulated comprehensive annotations for these documents. These annotations serve as essential training data for an AI model.

The team of annotators is heterogeneous, comprising individuals with diverse experiences and expertise. It consists of predominantly female individuals, all native Italian speakers. The annotators are all educated. Most of them have obtained their Ph.D. degrees or are currently pursuing them in linguistics. Fifty seven percent of them have at least 10 years of experience in linguistics, and 50% have at least 3 years of experience in inclusive language. Additionally, the annotators received an average of 30 hours of training on inclusive language and the annotation tool.

3.2 Linguistic Criteria

The process of defining linguistic criteria for inclusive communication involved the collaboration of the subset of linguistic experts with over 10 years of experience in inclusive writing. Initially, the team held discussions to share insights on inclusive language, which fostered open dialogue about its various aspects and challenges. After these discussions, the team drafted the linguistic criteria, incorporating each member's expertise. This writing phase included multiple revisions, during which the team evaluated the content for clarity and effectiveness. Further discussions allowed for addressing concerns and refining the linguistic criteria. Ultimately, the team reached a final decision, resulting in a well-structured set of criteria based on their expertise in linguistics, the Italian language, and culture. These criteria adeptly tackle issues related to gender, race, age, and disability while considering cross-category constraints. For instance, to overcome gender issues,

Table 2. Linguistic Criteria

Category	Description	*Non-inclusive* example	Reformulated *inclusive* example
Epicene	Use epicene words, gender-neutral words that can refer to people of any gender.	**IT:** *I corsi saranno tenuti dai professori.* **EN:** *The courses will be taught by professors.*	**IT:** *I corsi saranno tenuti da docenti a contratto.* **EN:** *Courses will be taught by contract faculty.*
Collective	Use collective nouns to refer to groups of people instead of specifying a gender.	**IT:** *Il professore deve preparare la lezione.* **EN:** *The professor must prepare the lesson.*	**IT:** *Il personale insegnante deve preparare la lezione.* **EN:** *Teaching staff must prepare the lesson.*
Metonymy	Use reformulations without an explicit agent obtained by metonymy.	**IT:** *Gli scienziati oggi hanno dimostrato la validità del modello.* **EN:** *Scientists today proved the validity of the model.*	**IT:** *La scienza oggi ha dimostrato la validità del modello.* **EN:** *Science today has proven the validity of the model.*
Generic	Use generic words like "person," "individual," or "citizen" instead of using gendered terms.	**IT:** *Tutti gli accompagnatori sono pregati di scendere.* **EN:** *All chaperones are requested to disembark.*	**IT:** *Tutte le persone che accompagnano sono pregate di scendere.* **EN:** *All accompanying people are requested to disembark.*
Omission	Omit non-inclusive segments when possible without affecting the meaning of the sentence.	**IT:** *Il professore può preparare la lezione.* **EN:** *The lecturer can prepare the lesson.*	**IT:** *È possibile preparare la lezione.* **EN:** *It is possible to prepare the lesson.*

Description and qualitative examples pertaining to specific reformulation rules devised by linguistic experts. For each criterion of inclusiveness, it reports the description, a *non-inclusive* example, and a corresponding possible *inclusive* reformulation produced by exploiting the criteria principle. For the sake of readability, the English translations are reported in gray.

using symbols like ə, *, or / was considered non-inclusive because it makes the reformulated text not readable by visually impaired people through automatic reading tools. Table 2 summarizes the main types of reformulation rules and provides qualitative examples:

— *Epicene*: Gendered words can be replaced by epicene ones, which are gender-neutral words that can refer to people of any gender, to increase the inclusivity of the sentence without changing the meaning. For instance, "*professori*" (EN: "*professors*") can be replaced with the epicene word "*docenti*" (EN: "*faculty*").
— *Collective*: It is preferable to use collective nouns instead of gender-specific nouns. For instance, instead of using the phrase "*il professore*" (Italian male declension of "*the professor*"), writers could use a collective noun such as "*il personale docente*" (EN: "*the teaching staff*"). In such a way, the sentence becomes inclusive because it no longer refers to a specific gender.
— *Metonymy*: Metonymy is a figure of speech that entails substituting the name of one thing for that of another, of which it is an attribute or with which it is associated (e.g., the "crown" is used to mean "king" or "queen"). Metonymy can increase inclusiveness by replacing explicit agents in a sentence. For instance, "*gli scienziati*" (EN: "*the scientists*"), which refers to a specific agent, can be replaced with "*la scienza*" (EN: "*science*").
— *Generic*: Another possibility is to replace gender-specific terms with similar generic words. For example, "*gli accompagnatori*" (EN: "*the chaperones*"), which in Italian is gender-specific, can be replaced by a generic term such as "*le persone che accompagnano*" (EN: "*the accompanying people*").

— *Omission*: Non-inclusive segments can be omitted if they do not affect the sentence's meaning. For example, in the sentence "*Il docente può preparare la lezione*" (EN: "*The lecturer can prepare the lesson*"), the non-inclusive segment that refers to an agent "*Il docente*" can be omitted without changing the meaning of the sentence. The omission produces "*È possibile preparare la lezione*" (EN: "*It is possible to prepare the lesson*").

3.3 Data Gathering and Annotation

Data Gathering. We collect documents written in Italian that reflect the language's usage in the administrative domain. To achieve our goal, we collect Italian documents from various sources, including the Web sites of public administrations and universities. Domain experts carefully select these documents to create a comprehensive dataset that captures a wide range of formal communication contexts. Specifically, we gather materials from public entities such as Città Metropolitana di Torino, Università di Bologna, Politecnico di Torino, and other reputable Italian institutions.

The collected documents represent various administrative genres and styles, including calls for bids, internal communications, policies, and regulations. The collected documents are split and annotated at the sentence level according to linguistic criteria for non-inclusive language detection and reformulation tasks.

Non-Inclusive Language Detection Annotation. We define the detection task as a *multi-class* sentence classification problem with mutually exclusive labels. Notice that the annotation is performed at the sentence level. However, multiple subjects or entities that can potentially contain inclusivity issues can be present within each sentence. Therefore, we annotated each sentence with one of the following labels:

— *Neutral*. The sentence does not contain any reference to a protected group. Therefore, it is not relevant to the task of discriminating between inclusive and non-inclusive language. However, these sentences are still relevant to the annotation process as they represent a sentence type likely to occur in real documents. Some examples of neutral sentences in formal documents also include openings and signatures.
— *Inclusive*. The sentence contains at least one explicit reference to a protected group (thus is relevant to the task), and *all* the related expressions are written inclusively. Therefore, if multiple inclusivity-related entities are present, they must be *all* written adopting inclusive expressions to consider the sentence as *inclusive*.
— *Non-Inclusive*. The sentence contains at least one reference to a protected group (thus is relevant to the task), and *at least one* of the related expressions is written in a non-inclusive manner. Therefore, if some expressions are written inclusively while others are non-inclusive, or all expressions are non-inclusive, the sentence is labeled as non-inclusive.

Table 3 reports one example per class and the corresponding English translation. The first sentence does not reference people or groups, which could harm inclusiveness. For this reason, it is labeled as *neutral*. The second sentence contains one potentially harmful expression (written in boldface). However, it has already been written in an inclusive way. Therefore, it is labeled as *inclusive* because *all* the potentially harmful expressions are written in inclusive language. Finally, the third sentence contains one potentially harmful expression (written in boldface), which is written in a non-inclusive manner. Therefore, the sentence is labeled as *non-inclusive*.

Inclusive Language Reformulation Annotation. The reformulation task is a sequence-to-sequence problem where input and output sentences belong to the same language. Therefore, annotators are asked to provide one or more possible inclusive reformulations of the *non-inclusive* sentences,

Table 3. Dataset Qualitative Examples

(1) Neutral
IT: *Gli eventi comprendono esperimenti e dimostrazioni scientifiche dal vivo.*
EN: *Events include live science experiments and demonstrations.*
(2) Inclusive
IT: *La scadenza del bando per* ***la componente studentesca*** *è il 20 dicembre.*
EN: *The call for applications deadline for* ***the student body*** *is 20 December.*
(3) Non-inclusive
IT: *La scadenza del bando per* ***gli studenti*** *è il 20 dicembre.*
EN: *The call for applications deadline for* ***the students*** *is 20 December.*

Examples of *neutral*, *inclusive*, and *non-inclusive* sentences in Italian (IT) and their corresponding translation in English (EN). The inclusive harmful expressions are written in boldface. Notice that in English, gender issues are not as evident as in the Italian language.

following the defined criteria. The proposed reformulation must inclusively rephrase all non-inclusive expressions while preserving the meaning and maintaining the correctness of the sentence. In this way, pairs of non-inclusive and inclusive formulations of sentences are created. In Table 3, the third sentence contains the non-inclusive expression "*gli studenti*" (EN: "*the students*"), which the expert has reformulated as in the second sentence by removing all gender references as "*la componente studentesca*" (EN: "*the student body*"). Hence, this pair of sentences (*non-inclusive* → *inclusive*) constitutes a single example within the reformulation dataset.

Annotation Process in Practice. Once the documents are selected, they are automatically processed to extract all text snippets. Each snippet is then split into sentences.[3] All extracted sentences are then uploaded to a customized annotation platform based on Label Studio [50]. For each sentence, the annotator must first select the inclusivity label from the set *neutral*, *inclusive*, and *non-inclusive*. Subsequently, for each *non-inclusive* sentence, the annotator proposes one or more *inclusive* reformulations. The annotation platform also allows annotators to suggest *non-inclusive* versions of already *inclusive* sentences to increase the number of reformulation pairs. However, this feature is currently used minimally as most source documents are written in non-inclusive language.[4] The final classification dataset comprises the distinct set of all neutral sentences, original non-inclusive sentences, and proposed inclusive reformulations. The reformulation dataset comprises all the pairs of non-inclusive sentences and inclusive reformulations. Consequently, the two datasets contain common sentences.

Template-Based Generated Data. The linguist experts also generated pairs of non-inclusive sentences with inclusive reformulations by exploiting templates. Each template contains a specific placeholder (e.g., `[FILL_ACTOR]`) that can be filled with both inclusive and non-inclusive versions. This approach allows for systematically exploring different reformulations that can mitigate non-inclusive language while preserving the original meaning. For instance, a sentence such as "`[FILL_ACTOR]` *parteciperà alla lezione*" (EN: "`[FILL_ACTOR]` *will attend the lecture*") can be reformulated as "`[Chi frequenta la scuola]` *parteciperà alla lezione*" (EN: "`[Those who attend the school]` *will attend the lecture*") in the inclusive version, and "`[L'alunno]` *parteciperà alla lezione*" (EN: "`[The student]` *will attend the lecture*") in the non-inclusive one. Template-generated sentences differ

[3]Utilizing https://spacy.io/models/it#it_core_news_sm.

[4]The annotation platform also supports additional annotations, such as named entity tags and labels, valuable for future research purposes [35].

Table 4. Dataset Characterization

Data split	# samples	# words	# hapax	Avg. word length
Detection dataset				
Inclusive	4,126	13,259	3,390	36.68
Non-inclusive	4,182	13,278	3,532	35.80
Neutral	2,492	6,396	2,250	13.68
Total	10,725	16,241	1,960	30.96
Reformulation dataset				
Inclusive	4,705	8,578	2,939	35.75
Non-inclusive	4,705	8,691	3,036	36.16

For the *detection* and *reformulation* datasets are reported the number of annotated sentences corresponding to each class and the total (# samples), the total number of distinct words (# words), the number of words for which there is only one occurrence (# hapax), and the mean number of words in those sentences (avg. word length).

significantly from those in [4]. They have been generated based on the expertise acquired during the preliminary experiments. Specifically, in [4], the objective of template-generated sentences was to produce the main training corpus. Instead, in this article, we exploit templates to generate types of sentences that are not frequent in the corpus but are important for the model to learn specific patterns of inclusiveness.

Ethical Considerations. All the collected data are publicly available on the Internet and free from copyright restrictions. Privacy and security concerns have been carefully considered throughout the data collection and annotation process by replacing references to proper names of people and institutions with randomly generated names to ensure privacy.

3.4 Data Characterization

The result of the data collection and annotation encompasses the following:

—227 official documents collected from different sources.
—10,725 distinct sentences labeled as *neutral*, *inclusive*, or *non-inclusive*.
—4,705 distinct pairs of *non-inclusive* sentence and *inclusive* text reformulation.
—75 template-based generated pairs of *non-inclusive* and *inclusive* sentences.

Table 4 shows the per-class and overall statistics about the collected dataset of Italian administrative documents. Specifically, it reports the number of samples annotated (*# samples*), the distinct number of words (*# words*), the number of words that only appear once (*# hapax*), and the mean number of words per sentence (*Avg. word length*).

The dataset contains approximately 16k distinct words, thus showing a fair variability in the collected documents. *Neutral* sentences are, on average, shorter than *inclusive* and *non-inclusive* ones. The reason is that administrative documents likely include short formulaic phrases such as headings, dates, formal openings, and signatures, which are inherently neutral. In contrast, *inclusive* and *non-inclusive* sentences retain a higher complexity and nuance in their content. Notice that *neutral* sentences are the minority class in the dataset, whereas 38% of the sentences contain non-inclusive forms.

4 Human-Validation Design

In this section, we describe the human validation tasks we designed to evaluate the outcomes of the detection (Section 4.1) and reformulation (Section 4.2) models. Specifically, we explain the guidelines we provided to ensure consistent and accurate annotation by linguistic experts and the labels we used to evaluate the quality of the proposed outcomes. Through this process, we aimed to obtain a comprehensive and accurate evaluation of the quality of the output generated by our models and identify areas for improvement and future research.

4.1 Human-Validation Task for the Detection Models

Non-inclusive language detection is a text classification problem (as introduced in Section 3). As we will report later in our analysis, we trained deep learning-based text classification models for this task (see Section 5.1). Despite the high performance obtained by deep learning models in text classification tasks, they behave as black-boxes [21, 27]. The opaque nature of such models hides the reasons behind their predictions. Even if they reach high performance in quantitative metrics such as accuracy and F1 score, their outcomes can be influenced by wrongly learned word patterns, model bias, or overfitting certain words. For instance, a sentence can be correctly classified for the right reason (i.e., using the words that determine that class) or for the wrong reason (i.e., using or overfitting other words that do not determine the class). Therefore, we designed a human validation task to ascertain if the classifiers use the correct set of words for their predictions. To do so, we first leverage state-of-the-art explainable AI techniques for NLP [9] to identify the most important words used by the models for the classification. Specifically, we apply *post-hoc* feature-based explainability techniques to the trained models and measure the importance of each input word in predicting the label for a given input sentence—we provide local explanations. Then, we ask inclusive language expert annotators to check whether the most important words used by the model to accomplish the classification task are the ones that correctly determine the prediction of the class label. Specifically, we design a survey where, for each sentence with the relative explanation (i.e., the most important words within each prediction highlighted in the text), the expert linguistic annotators should select one of the following options:

- *Correct (C).* The prediction is influenced by *only the words* determining the inclusiveness or the non-inclusiveness of the sentence, and all the entities present in the input sentences influenced the prediction.
- *Partially Correct (P).* The prediction is influenced by mainly words relevant to the concept of inclusiveness, but also includes a significant set of words or phrases that are not relevant for classification in terms of inclusiveness, or the production is influenced by a subset of the entities.
- *Incorrect (I).* The prediction is influenced by *only wrong words* that do not determine the inclusiveness or non-inclusiveness of the sentence, or all entities present in the sentence not influenced the prediction.

The evaluation process described above allows us to determine whether the classifier is using the correct set of words for the prediction (more details in Section 5.1).

4.2 Human-Validation Task for the Reformulation Models

Inclusive language reformulation is a sequence-to-sequence problem (as introduced in Section 3). Therefore, we trained deep learning text generation models for this task (as we will discuss in Section 5.1). While quantitative metrics such as BLEU [30] and ROUGE [26] can provide valuable insights into the performance of deep learning models in text generation tasks such as reformulation,

they are not always sufficient for evaluating the quality of NLP tasks, particularly those involving complex linguistic phenomena such as inclusivity. To address this limitation, we designed a human evaluation task in which linguistic expert annotators are asked to assign qualitative labels to each proposed reformulation:

— *Correct (C)*. The reformulation is correct and can be accepted without any further modification. The annotators are asked to consider the reformulation correct only if *all* the non-inclusive expressions are correctly reformulated, the grammar is correct, and the original meaning is preserved.
— *Partially Correct—Meaning Changed (P – M)*. The reformulation is not fully correct and requires human intervention to be accepted because the model partially changed the meaning of the original sentence while solving all the non-inclusive expressions.
— *Partially Correct—Subset (P – S)*. The reformulation is not fully correct and requires human intervention to be accepted because the model solved only a subset of the non-inclusive expressions.
— *Incorrect (I)*. The reformulation is incorrect because it does not solve any of the non-inclusive expressions.

This approach allowed us to obtain a more nuanced and comprehensive evaluation of the quality of the proposed reformulations, taking into account the complex linguistic phenomena involved in the task. Moreover, the two partially correct categories allow us to judge if the proposed reformulations are fluent and adequately preserve the original meaning of the sentence while addressing the non-inclusivity issues [44].

5 Model Learning and Evaluation

We train and evaluate deep learning models for non-inclusive language detection (Section 5.1) and reformulation (Section 5.2) tasks.

5.1 Non-Inclusive Language Detection

Model Learning. For each sentence, the detection model takes the sequence of words as input and predicts the most probable class label from the set: *neutral*, *inclusive*, or *non-inclusive* as output. We expected that the classification task would require advanced text understanding capabilities since the model needs to understand the sentence's meaning and detect patterns that make it potentially not inclusive. Therefore, we use two pre-trained transformer-based language models on the Italian language (*BERT-base-multilingual*[5] and *BERT-base-Italian*[6]), which are fine-tuned using the collected dataset (introduced in Section 3). We also trained three simpler baseline models based on a combination of **Term Frequency–Inverse Document Frequency (TF-IDF)** with a **Multi-Layer Perception (MLP)**, **Support Vector Machine (SVM)**, and **Gradient Boosting (GB)** for comparison. The classification dataset is split into training, evaluation, and test sets using an 80-10-10% ratio, resulting in 8,580-1,072-1,072 samples.

Quantitative Evaluation. To evaluate the performance of the detection models, we use both the *overall accuracy*, which measures the percentage of correctly classified sentences, and the *intra-class accuracy*, which measures the percentage of correctly classified sentences for each class. This allows us to evaluate the model's performance on the different classes, which may not be equally represented in the dataset.

[5]Original model available at: https://huggingface.co/bert-base-multilingual-cased.
[6]Original model available at: https://huggingface.co/dbmdz/bert-base-italian-cased.

Table 5. Quantitative Evaluation of the Detection Models

Model	Accuracy	Inclusive			Non-inclusive			Neutral		
		P	R	F1	P	R	F1	P	R	F1
TF-IDF + MLP	0.68	0.62	0.63	0.63	0.68	0.70	0.69	0.78	0.74	0.76
TF-IDF + SVM	0.61	0.53	0.53	0.53	0.56	0.63	0.60	**0.88**	0.70	0.78
TF-IDF + GB	0.74	0.74	0.74	0.74	0.76	0.76	0.76	0.71	0.72	0.72
BERT-multilingual	0.86	0.82	0.85	0.83	0.89	0.90	0.89	0.86	0.80	0.83
BERT-base-Italian	**0.89**	**0.88**	**0.88**	**0.88**	**0.92**	**0.92**	**0.92**	0.85	**0.86**	**0.85**

For each model are reported the overall *accuracy*, and the *precision* (P), the *recall* (R), and *F1 score* (F1) separately for each class label. The best-performing approach for each metric and the best model overall are highlighted in boldface.

Table 5 shows the classification performance in terms of *accuracy*, *precision* (P), *recall* (R), and *F1 score* (F1) for all the evaluated models. As expected, transformer-based models (*BERT-base-multilingual* and *BERT-base-Italian*) perform better on all evaluation metrics than the simpler baseline models. These models are able to distinguish between *neutral*, *inclusive*, and *non-inclusive* sentences in 86% and 89% of the cases, respectively. This shows that the task requires complex language understanding capabilities to be properly addressed. Interestingly, the Italian model performs better than the multi-lingual one. The lower performance of the multi-lingual model is probably because it is pre-trained on a mix of 100 different languages and may not appropriately capture the linguistic nuances required for precise classification in Italian. Notably, the Italian model is highly effective in detecting *non-inclusive* sentences by achieving a 0.92 F1 score and recall for the *non-inclusive* label.

Human Evaluation. For the human evaluation of the detection models, we conducted the designed human task described in Section 4.1. We evaluated only the two best-performing models (*BERT-Italian* and *BERT-multilingual*) because they achieved much higher performance than the simpler baseline. We randomly selected 100 correctly classified sentences with high confidence by both classifiers (with a probability higher than 0.90%) from the test set (50 *inclusive* and 50 *non-inclusive*). The *neutral* sentences do not contain specific patterns that must be learned by the models and are, therefore, not interesting to validate with such a human study. We aim to ascertain whether each sentence has been correctly classified for the correct or wrong reason (i.e., we analyze the most important/influential words used by the classifier that determine the predicted class label).

To produce the *local explanations*, we exploited T-EBAnO [55], the text explainer of the EBAnO framework [55, 56]. Each *local explanation* identifies the most important words influencing the model's prediction given an input sentence. Specifically, the *local explanations* produced by T-EBAnO identify the smallest set of words that causes the higher decrease of probability (or a change in the predicted class) when removed from the original input text. We chose T-EBAnO for the simplicity of the explanations produced and because it computes the importance of entire words instead of sub-words, making it more suitable for this evaluation since inclusivity expressions comprise one or multiple full words. Table 6 shows three examples of local explanations produced by T-EBAnO and evaluated in the human study. The classifier predicts the first sentence as *inclusive*; thereby, the highlighted words are the most important ones to determine the inclusiveness of the sentence for the model. In this case, the model focused on both inclusive expressions. According to Section 4.1, the classification and explanation can be considered *correct* because the prediction is influenced by only the words determining the inclusiveness of the sentence. Similarly, the second example shows a sentence correctly predicted as *non-inclusive*, where the classifier focused on all

Table 6. Examples of Local-Explanations Produced Using [55]

Sentence	Predicted label
IT: *La comunicazione è rivolta alla componente studentesca e al personale dipendente dell'università.* **EN:** *The communication is addressed to the student body and the staff of the university.*	*Inclusive*
IT: *La comunicazione è rivolta a tutti gli studenti e ai dipendenti dell'università.* **EN:** *The communication is addressed to [MALE] all students and [MALE] university employees.*	*Non-inclusive*
IT: *Sarà più semplice aiutare studenti ed educatori a sviluppare gli strumenti per affrontare in modo aperto, rispettoso e sicuro per tutti le discussioni sui temi della differenza etnica e culturale.* **EN:** *It will be easier to help [MALE] students and [MALE] educators develop the tools to approach discussions on issues of ethnic and cultural difference in an open, respectful and safe manner for [MALE]all.*	*Non-inclusive*

The highlighted words represent the ones most influencing the predicted label by the model for each individual sentence. The first two explanations reveal that the classifier is influenced by all the words related to inclusivity; the third one is influenced by a subset of words instead.

the non-inclusive expressions. Therefore, this classification and explanation can also be considered *correct*. Finally, the third sentence is correctly classified as *non-inclusive*. However, the model was influenced by a partial subset of the non-inclusive expressions ("*studenti*" and "*educatori*"), but not by the other non-inclusive expression "*tutti*," which is the masculine declension of the English word "*all*." Therefore, this classification and explanation can be considered *partially correct*.

Seven domain experts served as annotators. Each annotator independently evaluated each sentence and explanation, selecting one of the labels defined in Section 4.1 (i.e., *correct*, *partially correct*, and *incorrect*). The final category was determined through majority voting, with the label receiving the highest number of votes selected as the final label.[7] Table 7 shows, for the two evaluated models, the percentage of explanations labeled as *correct* (*C*), *partially correct* (*P*), and *incorrect* (*I*) for the 50 *inclusive* and *non-inclusive* sentences separately, and overall, by considering all 100 sentences. The annotators considered the explanations provided by the *BERT-base-Italian* model as the most accurate in justifying predictions for both *inclusive* and *non-inclusive* sentences separately, as well as overall. This model's explanations have been identified as *correct* (*C*) in 76% of cases for *inclusive* sentences and 98% for *non-inclusive* sentences, resulting in an overall correctness rate of 87%. In comparison, the fine-tuned *BERT-multilingual* model was rated as *correct* in 58% of explanations for *inclusive* sentences and 90% for *non-inclusive* sentences, with an overall correctness rate of 74%. Although both models generally provide accurate explanations for their predictions, with the *incorrect* label never exceeding 2%, the *BERT-base-Italian* model shows a stronger focus on relevant content, mainly when predicting the inclusive class. We also measured the inter-annotator agreements using Fleiss' Kappa, which measures the level of agreement between two or more human annotators on a particular task. Overall, the agreement between the seven participants for both models is 0.45. Following Landis and Koch's scale [25] to classify the level of agreement, which ranges from slight to almost perfect agreement, the obtained scores correspond to a fair agreement between annotators. This level of agreement likely reflects the inherent subjectivity in determining the exact set of words that contribute to a valid explanation.

We can conclude that, from a human perspective, the *BERT-base-Italian* model provides the most accurate explanations for its predictions. These results align with the model's superior performance shown in the quantitative evaluation, supporting its selection as the overall best-performing

[7] In cases of ties, the lower evaluation category was assigned (e.g., if tied between *correct* and *partially correct*, the latter was assigned).

Table 7. Human Evaluation of the Best-Performing Detection Models

Model	Inclusive			Non-inclusive			Overall		
	$C\uparrow$	$P\downarrow$	$I\downarrow$	$C\uparrow$	$P\downarrow$	$I\downarrow$	$C\uparrow$	$P\downarrow$	$I\downarrow$
BERT-multilingual	0.58	0.42	**0.00**	0.90	0.08	0.02	0.74	0.25	**0.01**
BERT-base-Italian	**0.76**	**0.22**	0.02	**0.98**	**0.02**	**0.00**	**0.87**	**0.12**	**0.01**

For both models are reported the percentage of explanations labeled as *correct* (*C*), *partially correct* (*P*), and *incorrect* (*I*) for the *inclusive* and *non-inclusive* sentences separately and overall. In bold, the best-performing values are shown for each metric.

model. Therefore, this classifier is currently integrated as the detection model of Inclusively (see Section 6.1).

5.2 Inclusive Language Reformulation

Model Learning. For the reformulation task, we fine-tuned the following T5-based pre-trained Italian models: (i) *IT5*[8] [40], and the efficient architecture (ii) *IT5-efficient*[9] [53]. We also fine-tuned two versions of the models by augmenting the training dataset with the 75 template-generated sentences defined by the inclusive language experts. The reformulation dataset is split into training, evaluation, and test sets using an 80-10-10 ratio, corresponding to 3,764-470-471 non-inclusive to inclusive pairs of reformulated sentences, respectively. For the models augmented with template-generated data, they were all added to the training set.

Quantitative Evaluation. To quantitatively assess the performance of the reformulation models, we exploited the BLEU [30] and ROUGE [26] scores. The former measures the similarity of individual reformulated sentences with a set of reference reformulated sentences. The latter instead measures the overlap n-grams. Specifically, we measured the precision (P), recall (R), and F1 score (R) of bigrams' overlap (ROUGE-2). All quantitative scores take into account neither intelligibility nor grammatical correctness.

Columns 3–6 in Table 8 show the performance in the evaluated quantitative metrics (BLEU and ROUGE-2) for the (i) *IT5* (×); (ii) *IT5-efficient* (×); (iii) *IT5 with templates* (✓); and (iv) *IT5-efficient with templates* (✓). Overall, we observed that templates slightly improved both models for all the evaluated metrics. The improved performance of template-augmented models may also be attributed to the fact that the training set has been expanded by a significant 2% of samples for those models. However, no clear best model emerges from the quantitative validation. It is worth noting that while the quantitative metrics used in this study, such as BLEU and ROUGE, are widely used in machine translation tasks, they may not be ideal for evaluating the quality of models in the context of inclusivity reformulation. These metrics are primarily designed to evaluate the similarity between two sets of text and may not fully capture the complex linguistic phenomena involved in reformulation tasks. As a result, no clear best model emerged from the quantitative validation alone. This highlights the importance of complementing quantitative metrics with human validation, as discussed in the previous section, to obtain a more comprehensive and accurate evaluation of the quality of the output generated by machine learning models in this specific NLP task.

Human Evaluation. For the human evaluation of the reformulation models, we carried out the designed task described in Section 4.2. Since all reformulation models achieved comparable quantitative performance, we validated all of them through human assessment. We randomly selected 100

[8]Original model available at: https://huggingface.co/gsarti/it5-base.
[9]Original model available at: https://huggingface.co/stefan-it/it5-efficient-small-el32.

Table 8. Quantitative and Human Evaluation of the Reformulation Models

Model	Templates	BLEU	ROUGE-2			Human evaluation			
			P	R	F1	C	$P-M\downarrow$	$P-S\downarrow$	$I\downarrow$
IT5-efficient	×	81.04	87.35	82.57	86.94	69.00	12.00	5.00	14.00
IT5-efficient	✓	**81.1**	87.47	**82.88**	87.05	73.00	13.00	3.00	11.00
IT5	×	80.32	87.42	82.3	87.17	68.00	13.00	7.00	12.00
IT5	✓	80.79	**87.68**	82.77	**87.47**	73.00	13.00	8.00	6.00

For each model, the BLEU, the ROUGE-2, and the human scores are reported. For the ROUGE-2, the *precision* (P), the *recall* (R), and the *F1 score* (F1) are reported. For the human validation are reported the percentage of reformulations labeled as *correct* (C), *partially correct—meaning changed* ($P-M$), *partially correct—subset* ($p-s$), and *incorrect* (I). In ✓are indicated the models trained with the template-based data augmentation. The best-performing approach for each metric and the best model overall are highlighted in boldface.

non-inclusive sentences from the test set, along with the corresponding reformulations generated by each model. Six expert annotators specializing in inclusive language were then asked to evaluate and label the proposed reformulations (accordingly to Section 4.2). Columns under *Human Evaluation* in Table 8 show the percentage of reformulations considered *correct* (C), *partially correct—meaning changed* ($P-M$), *partially correct—subset* ($P-S$), and *incorrect* (I) for all the reformulation models evaluated. Both IT5-efficient and IT5 models show similar performance, with templates consistently enhancing results for each. In both cases, template-based models achieve a *correct* (C) classification in 73% of instances, and the rate of *partially correct—meaning changed* ($P-M$) reformulations remains consistent across models. However, a notable difference appears in the distribution of other errors. The IT5-efficient model has a higher percentage of reformulations labeled as completely *incorrect* (I), whereas the IT5 model places more cases into the *partially correct—subset* ($P-S$) category. This distribution indicates that, even without achieving full correctness, the IT5 model can at least partially address inclusivity issues, thereby reducing the frequency of fully incorrect reformulations. Consequently, the IT5 model with templates may be preferable for its ability to make partial improvements more reliably.

Also, in this case, we measured the inter-annotator agreement using Fleiss's Kappa. The agreement for these tasks was 0.702, indicating a substantial agreement between human annotators. It is important to note that while the reformulation task can be inherently subjective, this level of agreement suggests a strong consensus on the quality of the reformulations. Dividing the *partially correct* cases into *meaning changed* ($P-M$) and *subset* ($P-S$) categories may have helped to improve consistency in evaluations.

Overall, we can conclude that all models achieved substantial levels of agreement between human annotators for the task of suggesting inclusive reformulation. The *IT5 model augmented with templates* proved to be the most reliable, achieving comparable or better results. Notably, the IT5 model showed fewer fully *incorrect* (I) cases, with more being classified as *partially correct—subset* ($P-S$), indicating more consistent partial corrections. This distribution reflects the IT5 model's reliability in generating reformulations that require minimal further intervention. Thus, the IT5 model is integrated as the reformulation model of Inclusively (see Section 6.2).

6 Inclusive Writing Assistance Tool: Inclusively

In this section, we present and summarize the main features of the writing assistance tool, namely Inclusively [24]. It acts as a text proofreader and a self-learning tool for non-expert writers. The tool is composed of an NLP pipeline that addresses two main tasks: (1) *detecting* non-inclusive harmful



Fig. 2. *Inclusive language modeling pipeline.* Sketch of the two-stage inclusive language modeling in the Inclusively tool for a single sentence. Firstly, the sentence is classified as *inclusive*, *non-inclusive*, or *neutral* by the detection model. Then, if the sentence is classified as *non-inclusive*, the reformulation model proposes a more inclusive reformulation. The procedure is the same for each sentence in a formal input document.

expressions (Section 6.1), and (2) *suggesting inclusive reformulations* for such expressions (Section 6.2). For the two tasks, Inclusively integrates the best-performing models based on the empirical evidence summarized in Section 5. Figure 2 exemplifies the presented pipeline: An Italian sentence is first classified as *non-inclusive* and then revised to produce an alternative, inclusive version.

6.1 Non-Inclusive Language Detection

The first stage of the Inclusively pipeline aims to highlight the snippets of text in the input documents that deserve attention due to the presence of inclusivity issues. Specifically, it takes as input a textual document, splits its content into sentences, and automatically detects non-inclusive portions of text within each sentence.

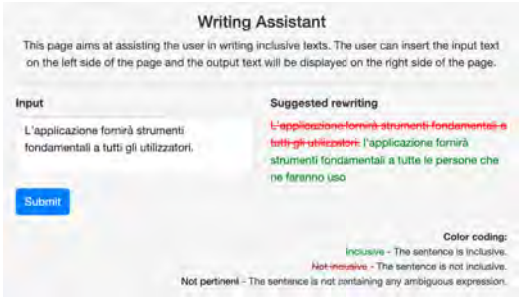
Every sentence is labeled as *inclusive*, *non-inclusive*, or *neutral* based on the categorization described in Section 3. Since each sentence belongs to exactly one category, the classification problem we address is *single-label* and *multi-class*.

The first stage of the pipeline currently integrates the fine-tuned *BERT-base-Italian* model, which is shown to be the best-performing based on the empirical results in Section 5.1. This model is able to distinguish between *neutral*, *inclusive*, and *non-inclusive* for approximately 9 sentences over 10. Notably, this model is particularly effective in the classification of *non-inclusive sentences* (it achieves 0.92 F1 score and recall for the *non-inclusive* label, as shown in Section 5.1). Correctly detecting *non-inclusive* sentences is crucial in the first stage of the pipeline.

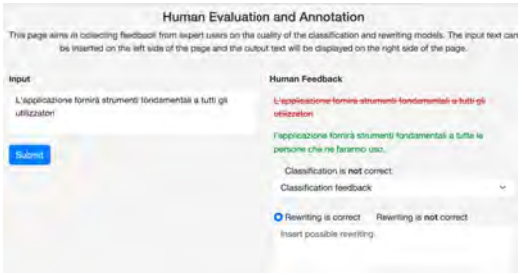
6.2 Inclusive Language Reformulation

The second stage of the Inclusively pipeline assists end-users in rewriting non-inclusive content. For all the sentences classified as *non-inclusive* by the detection model, the reformulation model should identify all the potentially non-inclusive expressions and rephrase each of them. The key goal is to mitigate the non-inclusivity issues while preserving the original meaning. Notice that while the detection task requires text understanding capabilities, the rephrasing task also requires text generation capabilities to rephrase the original content in a more inclusive way.

The second stage of the pipeline currently integrates the fine-tuned *IT5 model augmented with templates*, which is shown to be the best-performing based on the empirical results in Section 5.2. This model is able to produce entirely correct reformulations that mitigate all the non-inclusivity issues while preserving the original meaning and the grammatical correctness in 73% of the cases. These reformulations do not require any further intervention by the writer. Instead, for 22% of the sentences, this model suggests reformulations that require partial modification by the writer, while only 6% of the proposed reformulations are wrong or require substantial modification.



(a) Writing assistance interface



(b) Evaluation and annotation interface



(c) Inspection interface - classification

Fig. 3. *Examples of* Inclusively *interface.* The writing assistance interface (a) is designed to be used by standard users as a text proofreader for inclusive language. The evaluation and annotation interface (b) is intended for linguistic experts to provide feedback and new annotations. The inspection interface (c) integrates explainability techniques that allow data scientists to inspect the models' outputs and normal users to learn based on the models' explanations.

6.3 Tool's Interfaces

To make Inclusively accessible and user-friendly, we have developed a Web-based interface [24]. Specifically, the tool provides three different interfaces, each devoted to a particular group of users.

Writing Assistance Interface. The first interface is intended to support standard users such as non-expert writers to inclusive language writing. It exploits the pipeline composed of the detection and reformulation models. Users can paste their input text into a text box and click a button to process it. The interface highlights all non-inclusive sentences and proposes inclusive reformulations for each of them. Figure 3(a) shows an example of the use of such an interface. The input text introduced by the user contains a single sentence "*L'applicazione fornirà strumenti fondamentali a tutti gli utilizzatori*" (en: "*The application will provide basic tools for all users*") predicted as non-inclusive by the classifier, and is shown crossed out and colored in red. Then, the reformulation model suggests a more inclusive alternative to the input sentence, colored in green. Note that the example reports a short piece of text containing a single sentence just for the sake of readability. However, the system operates at the sentence level and can accommodate any number of sentences.

Evaluation and Annotation Interface. The second interface is designed for expert users, such as linguists or communication professionals, to provide feedback on the system's performance. Users can manually annotate sentences and model outcomes on this interface and provide feedback on the suggested reformulations. This feedback can then be used to evaluate and collect new data to fine-tune the system, ensuring its continued accuracy and relevance. Figure 3(b) shows an example of

the interface for the same input as the previous example. Users can indicate whether each sentence has been misclassified and propose alternative reformulations of the non-inclusive sentences.

Inspection Interface. The third interface is intended for data scientists. It includes explainability models [5, 39] that show which part of a sentence contributes most to predicting a specific label. This interface also highlights which words in the input sentence most contribute to the generated reformulations, allowing data scientists to understand better the system's behavior. Figure 3(c) shows the interface output to explain the detection model for the same input text. Users can identify the most important words the model uses for the classification. In this case, the words "*tutti*" (EN: "*all*") and "*utilizzatori*" (EN: "*users*") most influenced the model in predicting the sentence as non-inclusive. This interface provides data scientists with valuable insights into the inner workings of our AI-based tool, allowing them to better understand and improve the system's behavior. Additionally, this interface can serve non-expert users for self-training purposes, allowing them to analyze why a sentence is considered non-inclusive and how a reformulation was generated.

The INCLUSIVELY interfaces not only provide end-users with a user-friendly way to access the power of our AI-based tool for supporting inclusive writing but also serve as a means of promoting and improving such models. In fact, by providing a platform for expert end-users to provide feedback and annotations, we can continuously improve and fine-tune our system, ensuring its continued accuracy and effectiveness. Through this collaborative process, we hope to advance the state of the art in AI-based tools for promoting inclusivity in written communication.

The tool currently supports the Italian language, with short-term perspectives of extension to other Romance languages where inclusivity issues are prevailing (e.g., French, Spanish). A demo of the writing assistance tool is available online.[10]

7 Discussion and Future Work

In this article, in collaboration with linguistic experts, we established criteria for Italian inclusive language along with an annotated dataset. We then fine-tuned transformer-based models for non-inclusive language detection and inclusive language reformulation tasks. The quantitative and human-driven evaluation demonstrates the effectiveness of the trained models. The best-performing models have been employed in the two-stage pipeline of the writing assistance tool (INCLUSIVELY). The tool can foster inclusive communication by identifying sentences containing potentially harmful expressions and suggesting revised sentences using inclusive language. The proposed models can bring many benefits to inclusive writing in administration and academia and help increase diversity and inclusivity in our society.

The models developed in these experiments and the tool source code have been made publicly available[11] to foster the research community to use it with the goal of addressing the current limitations of state-of-the-art NLP models such as decision-making systems, language translators, and sentiment analyzers.

7.1 Limitations

While inclusive language is becoming an increasingly important aspect of written communication, and AI-based tools have shown promise in supporting the use of inclusive language, it is essential to acknowledge their limitations. Our current approach has four main limitations.

Italian Language-Specificity. The collected and annotated data, the trained models, and the developed application tool are specific to the Italian language. As a result, the language specificity of the

[10] https://youtu.be/3uiW_ti8wmY.
[11] https://github.com/MorenoLaQuatra/inclusively.

dataset and models inherently limits their applicability in multilingual or non-Italian contexts. Consequently, while this article demonstrates the potential of the detection and reformulation models for promoting inclusivity using expert-annotated data within the Italian language, the applicability and effectiveness of the approach, when applied to other languages without substantial re-training and re-annotation efforts. The approach's reliance on language-specific features underscores the challenge of generalizing these findings across different linguistic contexts, highlighting the need for further research and development to adapt similar methodologies to other languages.

Formal Communication-Specificity. Our study focuses on addressing the specific problem of inclusive language in administration and academic contexts. Therefore, the classification and reformulation models are trained on administrative documents, which may not be suitable for other contexts, such as legal and Web communications. Extending the approach to other communication domains requires not only annotation and training efforts but also the definition of appropriate linguistic criteria tailored to each specific data domain.

Dataset Size. Despite the effort to collect high-quality expert-annotated data for inclusive language in Italian, it may not cover all possible expressions encountered in formal communication. We acknowledge that the size of the dataset is relatively small. This limitation stems from the challenge of acquiring sufficient examples that accurately represent the diversity and nuance of inclusive language within formal communication. Consequently, while our dataset provides a valuable resource for studying inclusive language, it may not encompass the full range of expressions and variations that could be encountered in broader contexts. This limitation underscores the need for ongoing data collection and expansion to ensure broader coverage and robustness in future studies. However, our dataset is not a static resource; it is designed to evolve and improve over time through the interfaces provided in our inclusive language tool.

Improvable Reformulation Performance. Based on our analysis, the best reformulation model produces almost perfect reformulations (i.e., solving all the inclusivity issues while preserving the original meaning) in 73% of the cases, and partially correct reformulations in 21% of the cases. This indicates that while the model demonstrates a strong capability in addressing inclusivity, there is still room for improvement. The remaining 6% of cases highlight areas where the model either fails to fully address inclusivity issues or alters the original meaning to some extent. These shortcomings suggest the need for further refinement in the model's ability to understand context, semantic nuances, and the intricate balance required between inclusivity and meaning preservation. Future work will focus on enhancing the model's accuracy through the integration of more diverse training data and the development of advanced algorithms that better capture the subtleties of inclusive language reformulation. Additionally, user feedback mechanisms within our tool will play a critical role in iteratively improving the model's performance.

7.2 Future Work

To address these limitations, future work will focus on developing a multilingual tool that promotes inclusive communication across various domains. This will require additional training data and linguistic resources to ensure that the tool can identify and reformulate texts in multiple languages. One possible direction for future work is to extend the proposed methodology to new domains, such as legal and Web communications, and to develop models that are tailored to the specific linguistic features of these contexts. The tool's applicability will also be extended by including romance languages, such as French and Spanish, and low-resource languages. Finally, we want to spread the use of the tool on a large scale to assess its effectiveness. Through its interfaces, the

tool can also help us collect more data to improve the performance of the models (evaluation and annotation interface) and detect wrong behaviors (inspection interface).

References

[1] Bashar Alhafni, Nizar Habash, and Houda Bouamor. 2022. User-centric gender rewriting. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 618–631. DOI: https://doi.org/10.18653/v1/2022.naacl-main.46

[2] Chantal Amrhein, Florian Schottmann, Rico Sennrich, and Samuel Läubli. 2023. Exploiting biased models to de-bias text: A gender-fair rewriting model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* (*Vol. 1*: Long Papers). Association for Computational Linguistics, 4486–4506. . DOI: https://doi.org/10.18653/v1/2023.acl-long.246

[3] American Psychological Association. 2023. Inclusive Language Guidelines. Retrieved August 2023 from https://www.apa.org/about/apa/equity-diversity-inclusion/language-guidelines

[4] Giuseppe Attanasio, Salvatore Greco, Moreno La Quatra, Luca Cagliero, Michela Tonti, Tania Cerquitelli, and Rachele Raus. 2021. E-MIMIC: Empowering multilingual inclusive communication. In *Proceedings of the 2021 IEEE International Conference on Big Data (Big Data)*. IEEE, 4227–4234. DOI: https://doi.org/10.1109/BigData52589.2021.9671868

[5] Giuseppe Attanasio, Eliana Pastor, Chiara Di Bonaventura, and Debora Nozza. 2023. Ferret: A framework for benchmarking explainers on transformers. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations. Association for Computational Linguistics*.

[6] Tom Babinszki, Anna Cavender, Michael Gower, Jeffery Hoehl, Darcy Lima, Erich Manser, and Shari Trewin. 2019. Inclusive writing. In *Web Accessibility—A Foundation for Research* (2nd ed.). Yeliz Yesilada and Simon Harper (Eds.), Springer, 135–152. DOI: https://doi.org/10.1007/978-1-4471-7440-0_8

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Proceedings of the Advances in Neural Information Processing Systems*, Vol. 33, 1877–1901.

[8] Alessandra Teresa Cignarella, Mirko Lai, Marra Andrea, and Sanguinetti Manuela. 2021. La ministro è incinta: A twitter account of women's job titles in Italian. In *Proceedings of the 8th Italian Conference on Computational Linguistics (CLiC-it '21)*, Vol. 3033. CEUR-WS.org, 1–7.

[9] Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable AI for natural language processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, 447–459. Retrieved from https://aclanthology.org/2020.aacl-main.46

[10] Università degli studi dell'Aquila. 2023. Per un uso della lingua Italiana rispettoso dei generi. Retrieved August 2023 from https://www.univaq.it/include/utilities/blob.php?item=file&table=allegato&id=4925

[11] Università degli studi di Bologna. 2023. Linee guida per la visibilità del genere nella comunicazione istituzionale dell'università di Bologna. Retrieved August 2023 from https://www.unibo.it/it/allegati/linee-guida-per-la-visibilita-del-genere-nella-comunicazione-istituzionale-dell2019universita-di-bologna/@@download/file/Linee-Guida-Genere-2020.pdf

[12] Università degli studi di Milano. 2023. Vademecum sul linguaggio di genere. Retrieved August 2023 from https://www.unimi.it/sites/default/files/2021-12/Vademecumlinguaggio%20di%20genere_Universit%C3%A0%20degli%20Studi%20di%20Milano.pdf

[13] Università degli studi di Torino. 2023. un approccio di genere al linguaggio amministrativo. Retrieved August 2023 from https://www.unito.it/sites/default/files/linee_guida_approccio_genere.pdf

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1: Long and Short Papers). Association for Computational Linguistics, 4171–4186. DOI: https://doi.org/10.18653/v1/N19-1423

[15] Università Mediterranea di Reggio Calabria. 2023. Indicazioni per un uso del linguaggio rispettoso delle differenze. Retrieved August 2023 from https://www.unirc.it/documentazione/media/files/ateneo/pari_opportunita/File_allegato_2.pdf

[16] Università di Siena. 2023. Generi e linguaggi: Linee Guida per un Linguaggio administrative e istituzionale inclusivo. Retrieved August 2023 from https://www.unisi.it/comunicazione/linee-guida-un-linguaggio-amministrativo-e-istituzionale

[17] Politecnico di Torino. 2023. Guida pratica per una comunicazione inclusiva. Retrieved August 2023 from https://www.polito.it/sites/default/files/2023-07/vademecum%20esteso_170723.pdf

[18] Theodor Diesner-Mayer and Niels Seidel. 2022. Supporting gender-neutral writing in German. In *Proceedings of the Mensch Und Computer 2022*. ACM, New York, NY. DOI: https://doi.org/10.1145/3543758.3547566

[19] UK Government. 2014. Inclusive Communication. Retrieved August 2023 from https://www.gov.uk/government/publications/inclusive-communication

[20] UK Government. 2020. Breaking Down Gender Stereotypes in Legal Writing. Retrieved August 2023 from https://civilservice.blog.gov.uk/2020/01/10/breaking-down-gender-stereotypes-in-legal-writing/

[21] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM Computing Surveys* 51, 5, Article 93 (Aug. 2018), 42 pages. DOI: https://doi.org/10.1145/3236009

[22] Philip C. Kolin. 2007. *Successful Writing at Work*. Houghton Mifflin.

[23] Moreno La Quatra and Luca Cagliero. 2023. BART-IT: An efficient sequence-to-sequence model for Italian text summarization. *Future Internet* 15, 1 (2023), 15. DOI: https://doi.org/10.3390/fi15010015

[24] Moreno La Quatra, Salvatore Greco, Luca Cagliero, and Tania Cerquitelli. 2023. Inclusively: An AI-based assistant for inclusive writing. In *Machine Learning and Knowledge Discovery in Databases: Applied Data Science and Demo Track*. Gianmarco De Francisci Morales, Claudia Perlich, Natali Ruchansky, Nicolas Kourtellis, Elena Baralis, and Francesco Bonchi (Eds.), Springer Nature, Switzerland, Cham, 361–365. DOI: https://doi.org/10.1007/978-3-031-43430-3_31

[25] J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33, 1 (1977), 159–174.

[26] Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. Retrieved from https://aclanthology.org/W04-1013

[27] Haochen Liu, Yiqi Wang, Wenqi Fan, Xiaorui Liu, Yaxin Li, Shaili Jain, Yunhao Liu, Anil Jain, and Jiliang Tang. 2022. Trustworthy AI: A computational perspective. *ACM Transactions on Intelligent Systems and Technology* 14, 1 (Nov. 2022), 1–59. DOI: https://doi.org/10.1145/3546872

[28] Lorenzo De Mattei, Michele Cafagna, Felice Dell'Orletta, Malvina Nissim, and Marco Guerini. 2020. GePpeTto carves Italian into a language model. In *Proceedings of the 7th Italian Conference on Computational Linguistics (CLiC-it '20)*, Vol. 2769, CEUR Workshop Proceedings. CEUR-WS.org. Retrieved from http://ceur-ws.org/Vol-2769/paper_46.pdf

[29] United Nations. 2023. Inclusive Language Guidance. Retrieved August 2023 from https://equality.leeds.ac.uk/support-and-resources/inclusive-language-guidance/#:~:text=Inclusive%20language%20can%20help%20to,preferences%20about%20language%20and%20identity

[30] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 311–318. DOI: https://doi.org/10.3115/1073083.1073135

[31] European Parliament. 2018. GENDER-NEUTRAL LANGUAGE in the European Parliament. Retrieved August 2023 from https://www.europarl.europa.eu/cmsdata/151780/GNL_Guidelines_EN.pdf

[32] European Parliament. 2018. LA NEUTRALITÀ DI GENERE NELLINGUAGGIO usato al Parlamento europeo. Retrieved August 2023 from https://www.provincia.mantova.it/UploadDocs/7990_linee_guida_Parlamento_Europeo.pdf

[33] Andrea Piergentili, Dennis Fucci, Beatrice Savoldi, Luisa Bentivogli, and Matteo Negri. 2023. Gender neutralization for an inclusive machine translation: From theoretical foundations to open challenges. In *Proceedings of the 1st Workshop on Gender-Inclusive Translation Technologies*. European Association for Machine Translation, 71–83. Retrieved from https://aclanthology.org/2023.gitt-1.7

[34] Marco Polignano, Pierpaolo Basile, Marco De Gemmis, Giovanni Semeraro, and Valerio Basile. 2019. ALBERTO: Italian BERT language understanding model for NLP challenging tasks based on tweets. In *Proceedings of the 6th Italian Conference on Computational Linguistics (CLiC-it '19)*, Vol. 2481. CEUR, 1–6.

[35] Moreno La Quatra, Salvatore Greco, Luca Cagliero, Michela Tonti, Francesca Dragotto, Rachele Raus, Stefania Cavagnoli, and Tania Cerquitelli. 2024. Building foundations for inclusiveness through expert-annotated data. In *Proceedings of the EDBT/ICDT Workshops*. Retrieved from https://ceur-ws.org/Vol-3651/DARLI-AP-3.pdf

[36] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog* 1, 8 (2019), 9.

[37] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research* 21, 140 (2020), 1–67. Retrieved from http://jmlr.org/papers/v21/20-074.html

[38] Rachele Raus, Michela Tonti, Tania Cerquitelli, Luca Cagliero, Giuseppe Attanasio, Moreno La Quatra, and Salvatore Greco. 2022. L'analyse du discours et l'intelligence artificielle pour réaliser une écriture inclusive : le projet EMIMIC. *SHS Web of Conferences* 138 (2022), 01007.DOI: 01007

[39] Gabriele Sarti, Nils Feldhus, Ludwig Sickert, Oskar van der Wal, Malvina Nissim, and Arianna Bisazza. 2023. Inseq: An interpretability toolkit for sequence generation models. In *Proceedings of the 61st Annual Meeting of the Association*

for Computational Linguistics (Vol. 3: System Demonstrations). Association for Computational Linguistics, 421–435. DOI: https://doi.org/10.18653/v1/2023.acl-demo.40

[40] Gabriele Sarti and Malvina Nissim. 2024. IT5: Text-to-text pretraining for Italian language understanding and generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING '24)*. ELRA and ICCL, 9422–9433. Retrieved from https://aclanthology.org/2024.lrec-main.823/

[41] Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation. *Transactions of the Association for Computational Linguistics* 9, 8 (2021), 845–874. DOI: https://doi.org/10.1162/tacl_a_00401

[42] Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2022. Under the morphosyntactic lens: A multifaceted evaluation of gender bias in speech translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (Vol. 1: Long Papers). Association for Computational Linguistics, 1807–1824. DOI: https://doi.org/10.18653/v1/2022.acl-long.127

[43] Beatrice Savoldi, Sara Papi, Matteo Negri, Ana Guerberof-Arenas, and Luisa Bentivogli. 2024. What the harm? Quantifying the tangible impact of gender bias in machine translation with a human-centered study. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 18048–18076. DOI: https://doi.org/10.18653/v1/2024.emnlp-main.1002

[44] Beatrice Savoldi, Andrea Piergentili, Dennis Fucci, Matteo Negri, and Luisa Bentivogli. 2024. A prompt response to the demand for automatic gender-neutral translation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics* (Vol. 2: Short Papers). Yvette Graham and Matthew Purver (Eds.), Association for Computational Linguistics, 256–267. Retrieved from https://aclanthology.org/2024.eacl-short.23

[45] Stefan Schweter. 2020. *Italian BERT and ELECTRA Models*. DOI: https://doi.org/10.5281/zenodo.4263142

[46] Karolina Stanczak and Isabelle Augenstein. 2021. A survey on gender bias in natural language processing. arXiv:2112.14168. Retrieved from https://arxiv.org/abs/2112.14168

[47] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 1630–1640. DOI: https://doi.org/10.18653/v1/P19-1159

[48] Tony Sun, Kellie Webster, Apurva Shah, William Yang Wang, and Melvin Johnson. 2021. They, them, theirs: Rewriting with gender-neutral English. arXiv:2102.06788. Retrieved from https://arxiv.org/abs/2102.06788

[49] Vivian P. Ta, Ryan L. Boyd, Sarah Seraj, Anne Keller, Caroline Griffith, Alexia Loggarakis, and Lael Medema. 2022. An inclusive, real-world investigation of persuasion in language and verbal behavior. *Journal of Computational Social Science* 5, 1 (2022), 883–903. DOI: https://doi.org/10.1007/s42001-021-00153-5

[50] Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020–2022. Label Studio: Data Labeling Software. Retrieved from https://github.com/heartexlabs/label-studio

[51] Disability Rights UK. 2023. Social Model of Disability: Language. Retrieved August 2023 from https://www.disabilityrightsuk.org/social-model-disability-language

[52] Eva Vanmassenhove, Chris Emmery, and Dimitar Shterionov. 2021. NeuTral rewriter: A rule-based and neural approach to automatic rewriting into gender neutral alternatives. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 8940–8948. DOI: https://doi.org/10.18653/v1/2021.emnlp-main.704

[53] Yi Tay, Mostafa Dehghani, Jinfeng Rao, William Fedus, Samira Abnar, Hyung Won Chung, Sharan Narang, Dani Yogatama, Ashish Vaswani, and Donald Metzler. 2022. Scale efficiently: Insights from pretraining and finetuning transformers. In *International Conference on Learning Representations*. Retrieved from https://openreview.net/forum?id=f2OYVDyfIB

[54] Leonor Veloso, Luisa Coheur, and Rui Ribeiro. 2023. A rewriting approach for gender inclusivity in Portuguese. In *Findings of the Association for Computational Linguistics (EMNLP '23)*. Houda Bouamor, Juan Pino, and Kalika Bali (Eds.), Association for Computational Linguistics, 8747–8759. DOI: https://doi.org/10.18653/v1/2023.findings-emnlp.585

[55] Francesco Ventura, Salvatore Greco, Daniele Apiletti, and Tania Cerquitelli. 2022. Trusting deep learning natural-language models via local and global explanations. *Knowledge and Information Systems* 64, 7 (Jul. 2022), 1863–1907. DOI: https://doi.org/10.1007/s10115-022-01690-9

[56] Francesco Ventura, Salvatore Greco, Daniele Apiletti, and Tania Cerquitelli. 2023. Explaining deep convolutional models by measuring the influence of interpretable features in image classification. *Data Mining and Knowledge Discovery* 38, 5 (Feb. 2023), 3169–3226. DOI: https://doi.org/10.1007/s10618-023-00915-x

Received 4 October 2023; revised 26 February 2025; accepted 29 March 2025