E-MIMIC: Empowering Multilingual Inclusive Communication

E-MIMIC: Empowering Multilingual Inclusive Communication

Giuseppe Attanasio, Salvatore Greco[1], Moreno La Quatra,
Luca Cagliero, Michela Tonti, Tania Cerquitelli
Department of Control and Computer Engineering
Politecnico di Torino
Turin, Italy
name.surname@polito.it
[1]name_surname@polito.it

Rachele Raus
Department of Interpretation and Translation
Università di Bologna
Bologna, Italy
rachele.raus@unibo.it

***Abstract*—Preserving diversity and inclusion is becoming a compelling need in both industry and academia. The ability to use appropriate forms of writing, speaking, and gestures is not widespread even in formal communications such as public calls, public announcements, official reports, and legal documents. The improper use of linguistic expressions can foment unacceptable forms of exclusion, stereotypes as well as forms of verbal violence against minorities, including women. Furthermore, existing machine translation tools are not designed to generate inclusive content.**

The present paper investigates a joint effort of the research communities of linguistics and Deep Learning Natural Language Understanding in fighting against non-inclusive, prejudiced language forms. It presents a methodology aimed at tackling the improper use of language in formal communication, with a particular attention paid to Romanic languages (Italian, in particular). State-of-the-art Deep Language Modeling architectures are exploited to automatically identify non-inclusive text snippets, suggest alternative forms, and produce inclusive text rephrasing. A preliminary evaluation conducted on a benchmark dataset shows promising results, i.e., 85% accuracy in predicting inclusive/non-inclusive communications.

***Index Terms*—Inclusive Language, Gender Equality, Natural Language Processing, Deep Learning.**

I. Introduction

In recent years inclusive languages have received increasing attention from both the academic and industrial communities [1]. Inclusivity entails fighting against any discrimination conveyed by language understood in the broadest sense (i.e., language, images, gestures, etc.). With the goal of preserving diversity and inclusion, both academia and industry have sparked a great debate among the possible counteractions. For instance, targeted initiatives to eliminate stereotypes and forms of verbal violence against minorities have recently been proposed [2], [3]. Under this umbrella, a relevant effort has been devoted to addressing gender equality in linguistic terms, not only for women but also for other categories that the notion of gender implies (i.e., homosexuals, transgender) [4].

Data-driven tools learn from textual data how to compose well-structured text snippets by means of Deep Learning techniques. The diffusion of automated Machine Translation and conversational agent tools has made the problem of non-inclusive text generation even worse and widespread. As a matter of fact, since they commonly rely on English-written, non-gendered document corpora their capability to generate inclusive text is quite limited [5].

Formal communications entail the exchange of official information that flows along with the different levels of the organizational hierarchy and conforms to the prescribed professional rules, policy, standards, processes, and regulations of the organization. The modality of formal communications encompasses audio speeches (e.g., audios from conference calls and public announcements) and textual documents (e.g., official reports, legal documents, and public calls) [6]. Since the digitalization process has made a significant portion of them available in electronic form (e.g., pdf files, podcasts, Web pages), there is an increasing research interest in developing automated solutions to preserve inclusivity and diversity in formal communications [7]–[9].

The present paper describes the goals, methodology, and preliminary outcomes achieved by the *Empowering Multilingual Inclusive comMunICation* (E-MIMIC) project, which is a joint effort of linguistic and Data Science experts. It conjugates the wide experience of linguistic experts in recognizing and proofreading inappropriate forms of writing or speaking with the proficiency of data scientists experts of Deep Learning architectures to learn from data without explicit programming.

The main purpose of E-MIMIC is to investigate to what extent the most recent advances of Deep Learning and Natural Language Processing (NLP) can be helpful for fighting against prejudiced language forms, with particular attention paid to formal communications. To this aim, the problem of detecting and managing non-inclusive language forms in formal communications is formulated as a language bias detection and mitigation task. Although a huge body of prior work has already been presented in the literature, as discussed in [10], it is often unclear

(1) *How* to define what language "bias" means, e.g., what linguistic expressions are likely to be correlated with non-inclusive communications,

(2) *Whether* Deep Learning techniques are capable of debiasing the input text and producing an appropriate text rephrase.

(3) *To what extent* large-scale, general-purpose, multilingual collections (e.g., Wikipedia) are suitable for learning pre-trained models, which are conveniently fine-tuned for tackling specific tasks (e.g., text rephrasing and generation).

The present paper presents a methodology addressing the three above-mentioned research questions. Specifically, it describes an NLP pipeline that encompasses the collection of the raw data sources, the data preparation steps with the help of the domain experts, and the self-supervised language modeling phase pre-trained on general-purpose textual content and fine-tuned on expert-annotated data. Each model specialization is instrumental for a particular downstream task, i.e., classify text as *inclusive* or *not*, rephrase a sentence in an inclusive form, or generate new text adhering to linguistic expert's rules.

The paper also presents a preliminary validation of the text classification step on an Italian benchmark collection. The accuracy performance (i.e., 85% of correctly classified text snippets) confirms the direction of the ongoing project is promising. The envisioned approach will be extended to other kinds of communications and Romanic languages.

The paper is organized as follows. Sections II and III overview the prior work in the linguistics and machine learning areas, respectively. Specifically, Section II contextualizes the project in the current context of investigation, i.e., Italy's situation in terms of language inclusivity as a prime example of the main issues characterizing Romanic languages. Section III discusses and summarizes the state-of-the-art for deep learning and Natural Language Processing. Then, Sections IV and V respectively describe the motivations and the characteristics of the proposed method whereas Section VI shows the results of a preliminary case study of inclusive language classification exploiting a generated synthetic dataset. Finally, Section VII discusses the main open issues and the future research directions.

II. Challenges in Italian communication: a prime example

In Italy, prior works related to linguistic feminization and stereotypes[1] have repeatedly shown that the feminine, also supported by the mass media, continues to be silenced, and sexist or racist stereotypes are repeated. Recent studies [11]–[13] have shown how the images associated with Italian women in advertising reduce women to an object of sexual desire or the image of the mother, the linchpin of the Catholic family. However, this last type of stereotypical conceptualization associated with the "traditional" family is not limited to advertising or the media but pervades Italian discourse in general, especially institutional discourse. The same tendency seems to apply to other languages, neo-Latin or not. Recently, a meeting organized by Mondadori Università[2] highlighted the amalgam repeated by websites that inextricably links the image of blacks, as opposed to the image of white men, to weapons and thus indirectly to violence. Pictures and words conveyed through different discourses ultimately contribute to discriminating against specific categories and even to silence them, as in the case of women [14].

Language, with its semantic asymmetries and stereotypical words, contributes to these forms of exclusion and discrimination, especially in our time, as neural networks, often trained on these very large corpora when learning from discourses already marked by these forms of silence, discrimination, or verbal violence, repeat them once again and naturalize them on a large scale.

The author in [15] has spoken of *algorithmic discrimination* in this context. The latter has also been addressed in [16], [17]. Do a simple test with one of the most popular automatic translators (e.g., Google Translate, DeepL, and Reverso): if you ask for the translation of a French text declined in the feminine form, you will get an Italian text written in the masculine form. The main reasons for this behavior of the algorithm are:

- Multilingual automatic translators often switch from English to translating to and from other languages. However, English often uses epic words, words that apply to both the masculine and the feminine. These non-gendered words are then translated as masculine in the target language.
- Machine translation tools are mainly based on deep natural network models trained on authoritative data sources that meet the criteria recommended by [18]. Some examples of traditional data sources are generated by bilingual governments (e.g., Canada), international organizations (e.g., UN agencies, EU institutions...) that provide sufficiently large corpora to be useful in learning tasks [19]. Unfortunately, such data sources contain international linguistic variants of legal and/or political language, where the 'neutral' masculine is often preferred to denote categories.

In Italy, institutional websites and, in general, public administration privilege the usage of the masculine as a "neutral" form. However, in 2018 the Ministry of Education, Universities, and Research in Italy proposed a set of guidelines[3] for gender-inclusive language avoiding the spread of masculine syntagms allowing a uniquely male narrative (i.e., students' opinion, the researcher competitions, the researchers' night). Unfortunately, the effective exploitation of such guidelines is time-consuming since the number of available documents to be rewritten is very large and requires a lot of linguistic expertise to be implemented correctly. Furthermore, even when the masculine is used not to refer to generic categories but to precise gendered individuals, the words are not declined (the so-called "inclusive masculinity," see also [20]). We encourage the reader to look at the websites of some Italian universities (e.g., Rome La Sapienza, Bologna) wherein the personal section "who are you" menu, we find labels such as doctoral student, undergraduate, student, all declined in the masculine

[1]Osservatorio di Pavia, GLocal Media Monitoring Project (2015), https://www.osservatorio.it/download/GMMP_Italy.pdf (last access: November 2021)

[2]https://www.mondadorieducation.it/ (latest access: November 2021)

[3]https://www.miur.gov.it/-/linee-guida-per-l-uso-del-genere-nel-linguaggio-amministrativo-del-miur (latest access: November 2021)

also in case of female. A recent survey conducted on the sites of the University of Turin has shown precisely the presence of this exclusively male narration, as well as the fact that, even where they have begun to intervene with an inclusive language, the intervention has often been inconsistent at the level of a single document or too diversified at the level of multiple texts, if not downright erroneous, making the document not readable (e.g., due to the excessive use of "/" to mark the double masculine and feminine forms which make the text unreadable.

III. Deep learning for Natural Language Processing

The present paper leverages AI to automate the resolution of challenges due to non-inclusive communications such as those described in Section II. Unstructured data (e.g., text and images) is the main source for training AI-based systems. Most of those approaches are trained without human supervision to gain general knowledge from input data. Language Models (LMs) are Machine Learning solutions aimed at achieving a deep understanding of the natural language. They can be applied to both written and spoken languages and tailored to multilingual sources. LMs address a variety of Natural Language Processing tasks ranging from the assignment of predefined labels (i.e., the text classification task) to the automatic generation of new text. With the advent of Deep Learning (DL) architectures LMs have reached impressive quality scores according to both intrinsic and extrinsic evaluations [21].

Recurrent models [22] are established Neural Network models capable of processing arbitrary sequences of textual units (e.g., words) to produce either a single target (i.e., a many-to-one task) or a new sequence (many-to-many). Despite they have established as state-of-the-art approaches in LMs they have shown strong limitations in time and memory complexity due to their limited degree of parallelization [23]. In this regard, a relevant improvement has been achieved by transformer architectures [24], which leverage the attention mechanism to look for the token pairs in the sequence that mainly relate to each other. Transformers rely on an encoder-decoder architecture, which first produces a fixed-size vector representation of the input sequences and then exploits the encoded version to generate a new sequence.

State-of-the-art LMs take advantage of the transformer architecture to tackle NLP tasks. For example, BERT [25] is an established sentence encoder that performs self-supervised learning from large-scale collections of textual documents. It produces effective contextualized vector representations of sentences (consisting of up to 512 tokens) in the absence of human-annotated source data. Alternative approaches have addressed the encoding of longer text snippets (e.g., Longformer [26]), the use of transformer-based decoders to generate new text snippets (e.g., the GPT-based architectures [27]–[29]), and the combined use of encoder and decoder stacks to process an arbitrary input sequence and produce an appropriate output sequence (e.g., the Sequence-to-Sequence BART [30] and PEGASUS [31] models). Sequence-to-sequence models have been successfully applied to tackle complex NLP tasks such as Machine Translation, text rephrasing, and summarization. A summary of the most relevant prior works is given in Table I.

Traditional Neural Network models commonly require a large amount of training data to get reliable predictions. This would entail a considerable human effort in (semi-)automatic annotation of the analyzed data. Thanks to the adoption of a self-supervised approach, transformer-based LM architectures have enabled the unsupervised analysis of large-scale document collections in the model pretraining phase. For example, BERT [25] was trained on the entire Wikipedia and the Book Corpus dataset [32], whereas GPT-3 was trained on over 45 terabytes of Web and book-related data.

Pretrained LMs can be trained once and then reused to tackle various NLP tasks. The idea behind the aforesaid two-step process is to apply a fine-tuning step on top of the pre-trained model in order to adapt it to the new task. This leverages the general-purpose knowledge captured during the pre-training phase and specializes it using a significantly smaller amount of annotated data. This broadens the scope of the LMs to a variety of different tasks, ranging from Question Answering, text categorization, and Entity Recognition [33].

IV. Deep learning for inclusive communication

The project entitled *Empowering Multilingual Inclusive comMunICation* (E-MIMIC) aims at fostering inclusive communications in real-world scenarios. It provides end-users with an automated tool for textual document analysis focused on detecting and overcoming language inclusivity issues.

Currently, the developed solution is mainly focused on Italian documents coming from the academia and public administration domain. However, the generality of the presented methodology allows further extensions to languages other than Italian and to different application scenarios. More specifically, the main purpose of the E-MIMIC project is to overcome the discriminatory use of language within a text, both in terms of grammatical asymmetry (silencing of the feminine form) and semantic asymmetry due to the presence of stereotypes and further implementation of inclusive criteria towards other categories. To achieve this goal, E-MIMIC fosters the adoption of a Deep Learning-based methodology to process raw input text and identify discriminatory text snippets within an input text. A text classification module leverages the users' meta-linguistic reflexive abilities acquired through the analysis of a considerable amount of textual data. The data collection includes a large, general-purpose collection of unlabeled data suitable for self-supervised model pre-training and a smaller set of documents annotated by linguistic experts at the sentence level[4]. The classification model attends relevant portions of text that are most likely to be correlated with non-inclusive language forms. Thus, it is capable of contextualizing language

[4]For each sentence, the label indicates whether the sentence is formulated in an inclusive form or not.

TABLE I
MOST POPULAR TRANSFORMER-BASED ARCHITECTURES.

Publication	Type	Model
BERT [25]	Encoder	Transformer-based language encoder
LongFormer [26]	Encoder	Encoder for long sequences with local attention
GPT [27]–[29]	Decoder	Transformer-based decoder for language modeling
BART [30]	Encoder/Decoder	Sequence-to-sequence model for summarization and translation
Pegasus [31]	Encoder/Decoder	Sequence-to-sequence model for summarization

bias in the context of inclusive communication, i.e., it answers the question (1) posed in Section 1 (*How*).

E-MIMIC also suggests inline textual corrections by performing an intra-linguistic translation from the discriminatory form to the inclusive one. To this end, it is worth noticing that the use we make of the Italian language can be sexist, racist, or biased. Therefore, E-MIMIC calls for new approaches to de-biasing the language in a way that is pursuant to the required inclusivity standards, i.e., see question 2 (*Whether*) in Section 1.

Furthermore, E-MIMIC addresses the lack of domain-specific data by fostering the use of a pretrain-and-fine-tune paradigm currently adopted by state-of-the-art transformer-based encoders and decoders (e.g., [34]). Various large-scale, generic document corpora are currently available for unsupervised language modeling. Conversely, annotated textual data are not easy to retrieve, and, in particular, large annotation sets related to language inclusivity are, to the best of our knowledge, currently not available. Hence, we envision the adoption of the pretrain-and-fine-tune to accomplish not only the sentence classification task but also other, related generative steps such as text rephrasing and generation, i.e., see question 3 (*To what extent*) in Section 1. Notice that text generation can be selectively triggered only when the predicted inclusivity score is negative.

From a practical viewpoint, E-MIMIC fosters active collaborations between linguistic experts, who are in charge of annotating limited portions of data for model fine-tuning, and data scientists working in the NLP research domain. Amongst others, they aim at addressing the following research questions:

1) What linguistic criteria should be adopted within each application domain in order to effectively capture the underlying text dependencies using Deep Learning techniques?
2) How can we generalize such linguistic criteria in order to be inclusive for all minorities?
3) Which text portions are worth being rephrased due to the lack of language inclusivity?
4) What are the Deep Learning algorithms that are most effective in learning users' meta-linguistic reflexive abilities?
5) How can we best configure the deep learning algorithms?
6) How can we retrieve alternative expressions to replace discriminatory language forms?

We pose the aforesaid research questions to multidisciplinary teams consisting of both linguistic experts and data scientists, including experts of Deep NLP. Tackling these issues effectively and efficiently requires the complementary knowledge of technical and linguistic experts because advanced Deep Learning processes need to be supported by high-quality data annotations.

To the best of our knowledge, this project is the first attempt to empower inclusive formal communication with the goal of modeling users' meta-linguistic reflexive abilities. E-MIMIC pursues a tangible impact on modern society as it would be able to produce inclusive pieces of text, which in turn will be the seed for new inclusive language models tailored to both humans and machines.

V. PROPOSED METHODOLOGY

We describe here the main analytical steps envisioned for the E-MIMIC project. The project will entail the design and development of the NLP pipeline of steps depicted in Figure 1. It consists of:

- A *data collection* step, in which a large collection of documents is retrieved and collected into a unified repository.
- A *data labeling* step, in which some parts of the retrieved documents are manually annotated by linguistic experts.
- A *data modeling* step, in which Deep NLP models are pre-trained and fine-tuned for non-inclusive language detection, text rephrasing, and generation. The generated language models are able to capture various morphological, syntactical, and semantic properties of the inclusive language.

The E-MIMIC project has involved Italian annotators with a solid linguistic background to supervise the annotation process. As a next step, we plan to involve further external annotators from different countries as well.

A more thorough description of each step follows.

A. Data collection

Romanic languages such as Italian are particularly prone to non-inclusive phrasing. However, a limited amount of annotated data is available to train ad hoc large language models. To the best of our knowledge, there exists no publicly available dataset annotated for inclusive Italian language detection. Hence, performing accurate data annotations and analyses become of primary relevance for future research developments.

E-MIMIC mainly addresses the problem of inclusive language in administrative documents, grants, internal and external policies, and calls for applications. Universities' dedicated

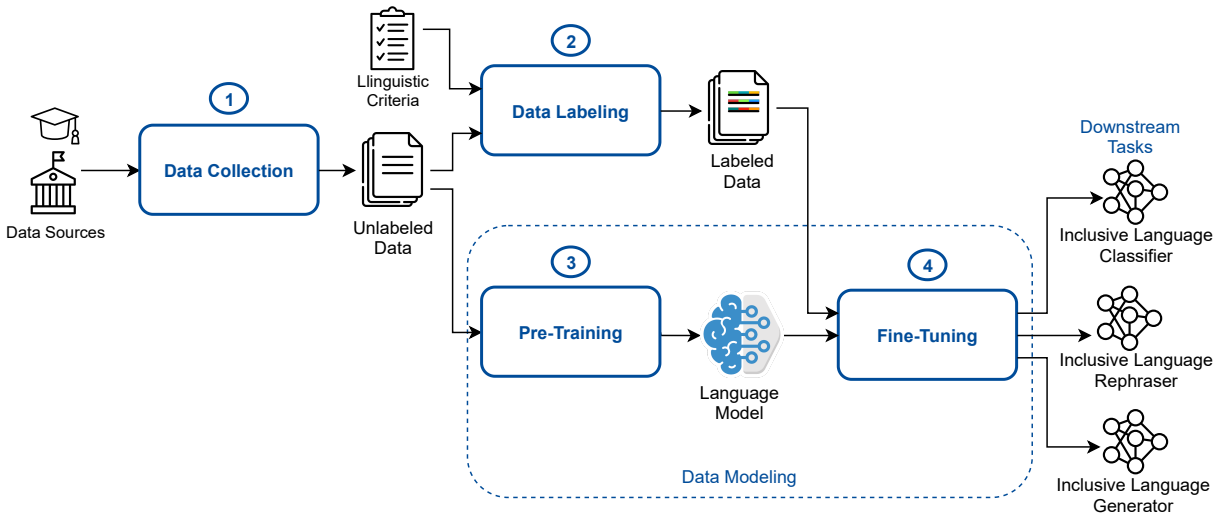

Fig. 1. E-MIMIC project. Sketch of the NLP pipeline.

offices and professionals commonly redact these kinds of documents for different stakeholders, namely students, faculties, public workers, university's employees, and its head roles. Although we focus on academic documents, the proposed approach can be easily extended to different types of documents, such as legal documents or official Webpages.

B. Data labeling

We envision building a comprehensive dataset for inclusive language. As such, we propose an annotation process spanning different facets of the problem. Given an arbitrary document, the annotation encompasses the following steps.

a) Sentence-level split: Before the actual annotation, we split each document into sentences. Therefore, we consider titles, headings, table cells, and every sentence in paragraphs as different data points. Although several other options exist (e.g., paragraph-based, word-based splitting), we consider sentence-level annotation because of the following reasons: (1) Words alone hardly convey non-inclusive concepts. (2) Sentences, instead, are the shortest unit to contain non-inclusive phrasing. (3) Sentences can easily be aggregated into paragraphs to account larger context [35].

b) Sentence annotation: We aim at training models that are able to extract inclusive and non-inclusive features from the raw text. As such, we both annotate linguist features and provide users with an alternative, inclusive re-formulation whenever a sentence contains non-inclusive phrases. Specifically, we annotate each sentence over the following aspects. 1) Inclusive label, either Inclusive, Non-Inclusive, or non-applicable. We consider a sentence non-inclusive if it contains at least one non-inclusive phrase or form. 2) Part-Of-Speech (POS) tagging of salient linguistic features, either cited content, proper names, and phrases potentially stereotypical. 3) Type of the content, either legal, administrative, technical, or informative. In case of non-inclusive phrasing, we 4) annotate the part of the sentence to be edited and 5) provide a re-phrasing of the sentence in its inclusive counterpart. To discount more than one valid re-formulation, we let annotators add multiple inclusive versions for the same sentence.

We rely on professional linguistics as annotators. Each annotator is trained on educational courses for inclusive language policies. Each document is annotated by one reviewer.

C. Modeling inclusive language

Due to the low resource settings, E-MIMIC leverages transfer learning, i.e., we build on off-the-shelf pre-trained language models. The latter model encodes lexical knowledge about the targeted language and hence serves as a solid starting point.

a) Target domain specialization: Firstly, we specialize models to language in the target domain (e.g., in our case, the administrative academic one). To this end, we envision the application of different specializing steps. First, a *Masked Language Model* (MLM) pre-training step can specialize the pre-existing language model to administrative language. After that, the resulting model is likely to be capable of understanding domain-specific word associations and hence can be used to generate coherent sentences. Part-Of-Speech (POS) tags can be optionally collected during the labeling phase to perform a *Named Entity Recognition* (NER). By doing so, the model learns how to contextualize based on the type of the processed words. Alternatively, models can be specialized via content classification, e.g., by predicting whether a sentence contains legal, administrative, technical, or informative content. Similar to named entity-based contextualization, models would learn language facets tailored to each application domain.

b) Learning inclusive writing: After domain-specific specializations, we envision a final training step, namely the fine-tuning on inclusive writing. At this stage, the model learns syntactical features and semantics of inclusive language based on the collected annotations. Two of the most common fine-tuning steps are inclusiveness classification and sequence-to-sequence learning.

Given the provided annotation, we can fine-tune language models as follows. First, we sample from annotated data sentences that are either inclusive or not. We then frame a binary classification task where models learn to distinguish inclusive sentences from non-inclusive ones. Users can apply such a model to scan long documents and highlight critical parts. Furthermore, we can leverage annotators' re-writing to fine-tune a sequence-to-sequence model. To accomplish the aforesaid task, the model learns how to align a sentence with its paired counterpart, e.g., in our context of analysis, a non-inclusive passage, and its inclusive re-writing. With high-quality learning, the resulting models can be used to re-write a non-inclusive sentence into a more inclusive version.

The two above-mentioned aspects can also be jointly addressed by a mixed solution. Given a whole document, a fine-tuned model can classify each sentence as *Inclusive* or *Non-Inclusive*. In the latter case, it can propose one or more possible reformulations.

VI. Case study

We carried out a preliminary empirical analysis of the effectiveness of the proposed methodology in a real case study. To this end, we built a language model able to detect whether a sentence follows inclusive writing criteria or not. The proposed study entails synthetic data generation, pre-trained model training, and fine-tuning for the classification task.

We used as new benchmark dataset collecting synthetically generated sentences in Italian. We generated the data samples using a template-filling procedure. Specifically, annotators formulated a *template* statement in the form of a sentence with a masked phrase to fill, where the masked portion denotes text inclusiveness. The top row in Table II reports a template instance. Next, we collect a parallel corpus of *seeds* used to fill templates' blanks. Each seed comprises two phrases, one following linguistic criteria for inclusive writing and one not. As such, the process of filling a template with an arbitrary seed generates two sentences, one inclusive and one non-inclusive. Bottom-most rows in Table II report an instance of the process. The inclusive example is denoted by *I*, whereas the non-inclusive one is denoted by *NI*.

We collected 19 templates and 43 seeds for a total of 822 samples.[5] Notice that, for each sentence, we also keep a binary label indicating whether it is inclusive or not.

We used the synthetic data to formulate a binary classification problem. Specifically, we randomly selected 80% of the templates to build the training set. The remaining ones compose the testing set. By restricting an arbitrary template to either the training or the testing set, we prevent the model from building only on word associations, avoiding context and linguistic criteria. Class labels were properly re-balanced in both sets.

[5]To preserve syntactic structure, we did not use every seed-template combination. Hence, the total number of generated samples was actually lower than their product.

TABLE II
Example of template-based synthetic data generation.

Template
Occorre richiedere la firma [blank]
Eng: *One must request the signature of [blank]*
Synthetic examples
NI: Occorre richiedere la firma degli interessati
I: Occorre richiere la firma delle persone interessate
Eng: *One must request the signature of interested people*

Concerning the classification model, we started testing a pre-trained BERT checkpoint[6] and fine-tuned it on the training data. We used the associated pre-trained sub-word tokenizer and kept 10% of the training as validation data.

The fine-tuned classifier achieved an accuracy of 85% on testing data (i.e., unseen templates). Although the preliminary outcomes were achieved on a synthetic, benchmark dataset, these results show the ability of modern neural architectures to effectively model inclusive language.

VII. Conclusions and discussion

In this paper, we presented the data analytics pipeline envisioned for the empowerment of a multilingual inclusive communication, whose main goal is to promote inclusive verbal communication (e.g., the institutional language level of public administrations). The problem is challenging but also urgent to build a better society by promoting inclusive communication.

Several stakeholders are involved in the development of the envisioned ML-based engine to effectively contribute their expertise to our multidisciplinary project. Linguists and data scientists are working together to share domain-specific knowledge and develop an expert system based on both deep learning models and linguistic criteria specific to each language. The development of the system will open various research issues to be addressed:

1) *Definition of a complete set of specific linguistic criteria*. Enumerating the relevant and complete set of linguistic criteria to be considered for each language is challenging due to the diversity of languages and scope (e.g., legal documents usually require different criteria than academic texts). We focus first on Italian, as it is one of the romance languages where the problem is most evident and can serve as a prime example of how promising and feasible the envisioned methodology is.
2) *Data Labeling.* This task is always time-consuming and crucial for the effectiveness of data-driven methods. High-quality standards, very large size, heterogeneous and diverse content are just some of the mandatory properties of the analyzed corpus. Moreover, different linguistics might prefer different linguistic forms to express inclusivity by increasing the complexity of different data-driven tasks.

[6]https://huggingface.co/dbmdz/bert-base-italian-cased (latest access: November 2021)

3) *Training of deep-learning accurate models.* This task relies on fine-tuning existing pre-trained models based on a very large and general corpus by exploiting ad hoc hardware resources. The task requires a high level of expertise in Deep Learning methods and an extensive knowledge of linguistic issues in order to find the best trade-off between the accuracy of the models and the time required to specialize the pre-trained models.
4) *Multi-faceted evaluation of deep-learning models.* The evaluation of deep-learning models is currently based on intrinsic and extrinsic quality metrics to be computed on a test set. The inherent complexity of the text classification and revision processes calls for new strategies to quantify the soundness and completeness of the achieved results.
5) *Tailoring the proposed methodology for different tasks.* The envisioned engine could be tailored to different tasks, using the acquired knowledge for simple tasks (e.g., classifying single sentences) to more complex tasks (classifying whole documents with a set of coherent recommendations to improve the inclusiveness of the overall documents).
6) *Personalized recommendations.* The proposed engine should suggest alternative texts to different users based on the application scenarios and users' knowledge and preferences. Appropriate strategies for collecting and analyzing the required data should be explored.

REFERENCES

[1] K. N. Canfield, S. Menezes, S. B. Matsuda, A. Moore, A. N. Mosley Austin, B. M. Dewsbury, M. I. Feliú-Mójer, K. W. B. McDuffie, K. Moore, C. A. Reich, H. M. Smith, and C. Taylor, "Science communication demands a critical approach that centers inclusion, equity, and intersectionality," *Frontiers in Communication*, vol. 5, p. 2, 2020.

[2] T. Redl, S. L. Frank, P. de Swart, and H. de Hoop, "The male bias of a generically-intended masculine pronoun: Evidence from eye-tracking and sentence evaluation," *PLOS ONE*, vol. 16, pp. 1–30, 04 2021.

[3] E. Laubscher, T. J. Raulston, and C. Ousley, "Supporting peer interactions in the inclusive preschool classroom using visual scene displays," *Journal of Special Education Technology*, vol. 0, no. 0, p. 0162643420981561, 2020.

[4] L. Zimman, "Transgender language reform: some challenges and strategies for promoting trans-affirming, gender-inclusive language," 2017.

[5] S. Jia, T. Meng, J. Zhao, and K. Chang, "Mitigating gender bias amplification in distribution by posterior regularization," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020* (D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, eds.), pp. 2936–2942, Association for Computational Linguistics, 2020.

[6] L. Qin, T. Xie, W. Che, and T. Liu, "A survey on spoken language understanding: Recent advances and new frontiers," in *IJCAI*, 2021.

[7] J. F. Tansey and E. S. Parks, "Privileged professionalisms: Using cocultural communication to strengthen inclusivity in professionalism education and community formation," *Ethics & Behavior*, vol. 0, no. 0, pp. 1–18, 2021.

[8] P. Millot, "Inclusivity and exclusivity in english as a business lingua franca: The expression of a professional voice in email communication," *English for Specific Purposes*, vol. 46, pp. 59–71, 2017.

[9] M. Varhelahti and T. Turnquist, "Diversity and communication in virtual project teams," *IEEE Transactions on Professional Communication*, vol. 64, no. 2, pp. 201–214, 2021.

[10] S. L. Blodgett, S. Barocas, H. D. III, and H. M. Wallach, "Language (technology) is power: A critical survey of "bias" in NLP," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020* (D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, eds.), pp. 5454–5476, Association for Computational Linguistics, 2020.

[11] P. Papakristo, *Il volto delle sirene. Storia della figura femminile nella pubblicità italiana*. Pesaro-Urbino: Aras edizioni, 2nd ed., 2018.

[12] M. Boero, *La famiglia della pubblicità. Stereotipi, ruoli, identità*. Milano: FrancoAngeli, 2nd ed., 2018.

[13] Gi.U.Li.A., *Stereotipi donne nei media*. Milano: LEdizioni, 2019.

[14] S. Jia, T. Lansdall-Welfare, and N. Cristianini, "Measuring gender bias in news images," in *Proceedings of the 24th International Conference on World Wide Web*, pp. 893–898, 2015.

[15] I. Bartoletti, *An Artificial Revolution: On Power, Politics and AI*. Mood Indigo, Indigo Press, 2020.

[16] E. Marzi, "La traduction automatique neuronale et les biais de genre : le cas des noms de métiers entre l'italien et le français," *Synergies Italie*, 2021.

[17] B. Savoldi, M. Gaido, L. Bentivogli, M. Negri, and M. Turchi, "Gender bias in machine translation," 04 2021.

[18] "ISO 31000:2009(en), Risk management — Principles and guidelines," Nov 2021. [Online; accessed 10. Nov. 2021].

[19] Y. Le Cun, *Quand la machine apprend: la révolution des neurones artificiels et de l'apprentissage profond*. Odile Jacob, 2019.

[20] R. Raus, "Per una cittadinanza della lingua: promuovere la parità di genere nel linguaggio amministrativo," 2016. [Online; accessed 10. Nov. 2021].

[21] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced BERT with Disentangled Attention," *arXiv*, Jun 2020.

[22] S. Sen and A. Raghunathan, "Approximate computing for long short term memory (LSTM) neural networks," *IEEE Trans. Comput. Aided Des. Integr. Circuits Syst.*, vol. 37, no. 11, pp. 2266–2276, 2018.

[23] J. Chung, Ç. Gülçehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *CoRR*, vol. abs/1412.3555, 2014.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017.

[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, June 2019.

[26] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *arXiv preprint arXiv:2004.05150*, 2020.

[27] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training,"

[28] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners,"

[29] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

[30] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (Online), pp. 7871–7880, Association for Computational Linguistics, July 2020.

[31] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization," in *International Conference on Machine Learning*, pp. 11328–11339, PMLR, 2020.

[32] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[33] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[34] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, June 2019.

[35] M. Yasunaga, J. Kasai, R. Zhang, A. Fabbri, I. Li, D. Friedman, and D. Radev, "ScisummNet: A large annotated corpus and content-impact

models for scientific paper summarization with citation networks," in *Proceedings of AAAI 2019*, 2019.