# Bias assessment for experts in discrimination, not in computer science

**Laura Alonso Alemany**[1,2], **Luciana Benotti**[1,2,3], **Hernán Maina**[1,2,3], **Lucía González**[1,2]
**Lautaro Martínez**[1,2], **Beatriz Busaniche**[2]
**Alexia Halvorsen**[2], **Amanda Mata Rojo**[2], **Mariela Rajngewerc**[1,2,3]
[1] Sección de Computación, FAMAF, Universidad Nacional de Córdoba
[2] Fundación Via Libre, Argentina
[3] Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina

## Abstract

Approaches to bias assessment usually require such technical skills that, by design, they leave discrimination experts out. In this paper we present EDIA, a tool that facilitates that experts in discrimination explore social biases in word embeddings and masked language models. Experts can then characterize those biases so that their presence can be assessed more systematically, and actions can be planned to address them. They can work interactively to assess the effects of different characterizations of bias in a given word embedding or language model, which helps to specify informal intuitions in concrete resources for systematic testing.

## 1 Introduction

Machine learning models and data-driven systems are increasingly being used to support decision-making processes. Such processes may affect fundamental rights, like the right to receive an education or the right to non-discrimination. It is important that models can be assessed and audited to guarantee that such rights are not compromised. Ideally, a wider range of actors should be able to carry out those audits, especially those that are knowledgeable of the context where systems are deployed or those that would be affected.

Several studies found that linguistic representations learned from corpora contain associations that produce harmful effects when brought into practice, like invisibilization, self-censorship or simply as deterrents (Blodgett et al., 2020). The effects of these associations on downstream applications have been treated as *bias*, that is, as *systematic errors* that affect some populations more than others, more than could be attributed to a random distribution of errors. This biased distribution of errors results in discrimination of those populations. Unsurprisingly, such discrimination often affects negatively populations that have been historically marginalized.

To detect and possibly reduce such harmful behaviour, many techniques for measuring and mitigating the bias encoded in word embeddings and Large Language Models (LLMs) have been proposed by NLP researchers and machine learning practitioners (Bolukbasi et al., 2016; Caliskan et al., 2017). In such works social scientists have been mainly reduced to the ancillary role of providing data for labeling, rather than being considered as core team (Kapoor and Narayanan, 2022). Current audits of data-driven systems often require technical skills that are beyond the capabilities of most of the people with knowledge on discrimination. The technical barrier has become a major hindrance to engaging experts and communities in the assessment of automated systems.

Moreover, we think approaching social risk mitigation through algorithmic calculations or adjustments is reductionist. We believe the part of the process that can most contribute to bias assessment are not subtle differences in metrics or technical complexities incrementally added to existing approaches, as is the case of a good portion of academic work in the area. Instead, we believe what can most contribute to an effective assessment of bias in NLP is precisely the linguistic characterization of the discrimination phenomena (Antoniak and Mimno, 2021).

That is why our aim with this work is to open up the participation of experts both on the complexities of the social world and on communities that are being directly affected by AI systems. Participation would allow processes to become transparent, accountable, and responsive to the needs of those directly affected by them.

The rest of this paper is organized as follows. In the next section we state the principles for integrating discrimination experts in the bias assessment process. We then review the shortcomings of



*Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 91–106

some approaches to bias assessment, and argue for the need for a tool specifically targeted to facilitate the integration of non-technical persons in the process of bias assessment. Then, we describe EDIA, the tool we developed to address this need. We finish with a discussion of our experiences in hands-on sessions with discrimination experts using the tool. Appendices with more extensive descriptions of the tool and a user story are also provided.

A demo of EDIA can be used at https://huggingface.co/spaces/vialibre/edia, allowing to explore the Word2Vec from Spanish Billion Word Corpus embedding (Cardellino, 2019) and BETO (Cañete et al., 2020) as the default language model. The tool is available at https://github.com/fvialibre/edia, and can be instantiated to explore different word embeddings and language models, independently of language, as is showcased in the Colab jupyter notebook illustrating the functionalities of EDIA[1].

## 2 Principles for integrating experts in discrimination in bias assessment

### 2.1 Interaction with discrimination experts to obtain an adequate tool

To create a tool that is truly useful for discrimination experts, we carried out hands-on workshops with diverse experts. In these workshops we declared our assumptions and motivations for the bias assessment process, observed their interaction with the methodology and obtained their feedback on the experience.

We carried out two workshops before the graphical interface was developed, then developed the interface integrating the requests and observations from those experiences and carried out two more workshops. We used a pre-survey and a post-survey to register the participants expertise and to record their experience with the tool, their suggestions for improvement and their requests for features. In particular we designed a questionnaire to collect the principles that they valued in the different versions of the prototype. During our workshops we also registered the workflow that the experts followed and we developed a user story that they reviewed that is published in (Benotti et al., 2023).

[1] https://colab.research.google.com/drive/1bSo9oXpB7fHjPB5UZGKJAcyA0zXHGjZO?usp=sharing

### 2.2 The principles

With our initial motivations and the insights gathered in these workshops, we developed EDIA, a tool for bias assessment in NLP artifacts, that follows the following design principles:

**Focus on expertise on discrimination,** substituting highly technical concepts by more intuitive concepts whenever possible and making technical complexities transparent in the process of exploration. More concretely, by hiding concepts like "*vector*", "*cosine*", etc. whenever possible, for example, substituting them for the more intuitive "*word*", "*contexts of occurrence*", "*similar*".

**Qualitative characterization of bias,** instead of metric-based diagnosis or mitigation.

**Integrate information about diverse aspects** of linguistic constructs and their contexts.

- provide context: which corpora, concrete contexts of occurrence (concordances), to get a more accurate idea of actual uses or meanings, even those that may have not been taken into account.
- provide information on statistical properties of words (mostly number of occurrences in the corpus, and relative frequency in different subcorpora), that may account for unsuspected behavior, like infrequent words being strongly associated to other words merely by chance occurrences.
- position with respect to other words in the embedding space, and most similar words.

**More complex representation of linguistic phenomena** word-based approaches are oversimplistic, and cannot deal with polysemy (the ambiguity or vagueness of words with respect to the meanings they may convey) or multiword expressions. That is why we need more context. Inspecting LLMs instead of word embeddings allows to account for those aspects of words. This has the added advantage of being able to inspect LLMs.

In designing these principles, we prioritized the specific needs of the Latin American region. In Latin America, we need domain experts to be able to carry out these analyses with autonomy, not relying on an interdisciplinary team or on training, since both are usually not available.

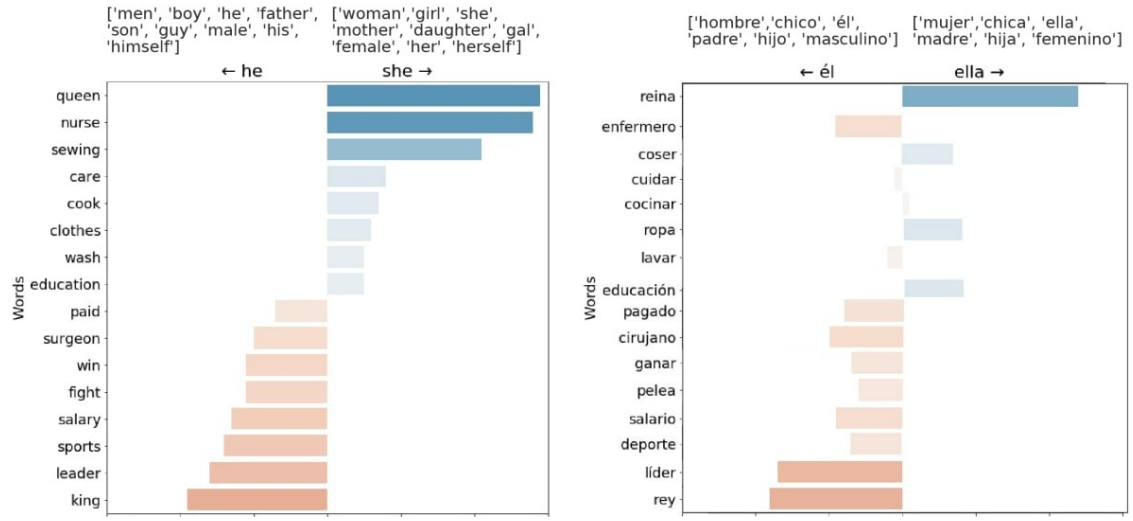


Figure 1: A list of 16 words in English (left) and a translation to Spanish (right) and the similarity of their word embeddings with respect to the list of words "*woman, girl, she, mother, daughter, feminine*" representing the concept "*feminine*", the list "*man, boy, he, father, son, masculine*" representing "*masculine*", and translations for both to Spanish. The English word embedding data and training is described in Bolukbasi et al. (2016) and the Spanish in by (Cañete et al., 2020). From the 16 words of interest, in English, 8 are more associated to the concept of "*feminine*", while in Spanish only 5 of them are. In particular, "*nurse*" in Spanish is morphologically marked with masculine gender in the word "*enfermero*" so, there is some degree of gender bias that needs to be taken into account to fully account for the behavior of the word. This figure illustrates that methodologies for bias detection developed for English are not directly applicable to other languages. Also, the figure illustrates that the observed biases depend completely on the list of words chosen.

## 3 A critical perspective on methods for bias assessment

In the last years the academic study of biases in language technology has been gaining growing relevance, with a variety of approaches accompanied by insightful critiques.

Early work on bias focused on finding metrics that allowed to adequately assess bias in word embeddings (i.e. Bolukbasi et al. (2016); Gonen and Goldberg (2019)). Most of the following work focused on technical subtleties about metrics, extensions to other languages or contexts, application to language models, evaluation on downstream tasks or automating the whole process, from assessment to mitigation (Guo and Caliskan, 2021; Guo et al., 2022; An et al., 2022; Kaneko and Bollegala, 2021).

### 3.1 On the importance of the linguistic representation of bias

Approaches to assess biases in word embeddings or large language models heavily rely on **lists of words** or **lists of sentences** to define the space of bias to be explored (Badilla et al., 2021). These resources have a crucial impact on how and which biases are detected and mitigated (Antoniak and Mimno, 2021), but they are not central in the efforts devoted to this task. The methodologies for choosing the words to make these lists are varied: sometimes lists are crowd-sourced, sometimes hand-selected by researchers, and sometimes drawn from prior work in the social sciences. Most of them are developed in one specific context and then used in others without reflection on the domain or context shift. They are even translated to other languages, disregarding linguistic and cultural differences that result in very different behaviors of the same word lists (Garg et al., 2018), as shown in Figure 1.

Most of the published work on biases exploration and mitigation has been produced by computer scientists based on the northern hemisphere, in big labs which have access to large amounts of funding, computing power and data. Unsurprisingly, most of the work has been carried out the English language and for gender and race biases (Garg et al., 2018; Blodgett et al., 2020; Field et al., 2021). Lauscher and Glavaš (2019) make a comparison on biases across different languages, embedding techniques, and texts. Zhou et al.

(2019) and Gonen et al. (2019) develop 2 different detection and mitigation techniques for languages with grammatical gender that are applied as a post processing technique. Even if they are targeting more diverse biases and languages, these approaches add many technical barriers that require extensive machine learning knowledge from the person that applies these techniques. Therefore they fail to engage interactively with relevant expertise outside the field of computer science, and with domain experts from particular NLP applications.

### 3.2 Criticisms to metric-centered approaches

Nissim et al. (2020) argue that the underlying assumptions for some of the metrics are inadequate. Jia et al. (2020) provide evidence that a reduction of bias shown in metrics does not correlate with a reduction of bias in downstream tasks. Even more worryingly, Antoniak and Mimno (2021) showed that metrics for bias assessment are very sensitive to changes in the word lists that are used as a basis for the diagnosis. They conclude that *word lists are probably unavoidable, but that no technical tool can absolve researchers from the duty to choose seeds carefully and intentionally.*

Blodgett et al. (2021) examine four sets of contrastive sentences to evaluate bias in language models and apply a method —originating from the social sciences— to inventory a range of pitfalls that threaten these benchmarks' validity as measurement models for stereotyping. They find that these benchmarks frequently lack clear articulations of what is being measured, and they highlight a range of ambiguities and unstated assumptions that affect how these benchmarks conceptualize and operationalize stereotyping. Névéol et al. (2022) propose how to overcome some of these challenges by taking a culturally aware standpoint and a curation methodology when designing such benchmarks.

With respect to mitigation, Brunet et al. (2019) show that debiasing techniques are more effective when applied to the texts wherefrom embeddings are induced, rather than applying them directly in the already induced word embeddings. Prost et al. (2019) show that overly simplistic mitigation strategies actually worsen fairness metrics in downstream tasks. More insightful mitigation strategies are required to actually debias the whole embedding and not only those words used to diagnose bias. However, debiasing input texts works best. Curating texts can be done automatically (Gonen et al., 2019) but this has yet to prove that it does not make matters worse. It is better that domain experts devise curation strategies for each particular case.

In spite of these well-founded critiques, work on bias in word embeddings and language models still revolves mainly around metrics and methods, and not so much on the participation of experts in the process of diagnosis. That is why we feel the need to facilitate the involvement of experts in bias assessment processes, so that the focus can be moved from technicalities to the problem itself.

In recent years, with the consolidation of bias assessment techniques, multiple frameworks have been developed to facilitate access to those techniques. We provide a description of some frameworks in Appendix B, and an overview of those with a graphical interface in Table 1.

Even in the case of those with a graphical interface, the design principles of those frameworks are still metric-centric, and most of them require mastery of machine learning methods and programming skills. Such requirements are usually barriers for non-technical profiles. As an alternative, we have developed EDIA, a no-code, no-statistics tool for experts to explore biases, which we describe in the following Section.

## 4 An intuitive tool to explore bias

This section provides a description of EDIA (acronym for the Spanish of *Stereotypes and Discrimination in Artificial Intelligence*), a visual interface framework for the analysis of bias in word embeddings and in LLMs[2]. A more detailed description of the tool can be seen in Benotti et al. (2023).

EDIA follows the design principles stated in 2, trying to fill a gap in the landscape of existing frameworks for bias assessment. It provides four main functionalities: exploring the learning data, exploring the distribution of words in an embedding space, systematizing biases in words and exploring biases in sentences. In what follows we describe these functionalities. In Appendix A we describe a user story showcasing how this tool may be used.

[2]EDIA is currently available at https://huggingface.co/spaces/vialibre/edia and https://github.com/fvialibre/edia.

| Framework | Reference | Word Embeddings Analysis | Language Models Analysis | Requieres NLP Knowledge | Mitigation Techniques Implemented | Counterfactuals Analysis |
|---|---|---|---|---|---|---|
| WordBias | Ghai et al. (2021) | ✓ | ✗ | ✗ | ✗ | ✗ |
| VERB | Rathore et al. (2021) | ✓ | ✗ | ✓ | ✓ | ✗ |
| LIT | Tenney et al. (2020) | ✓ | ✓ | ✓ | ✗ | ✓ |
| EDIA | Benotti et al. (2023) | ✓ | ✓ | ✗ | ✗ | ✗ |

Table 1: Description of frameworks with graphical interfaces available for bias analysis of embeddings or language models. The What-if Tool is not included in the table because it does not specifically target text data.

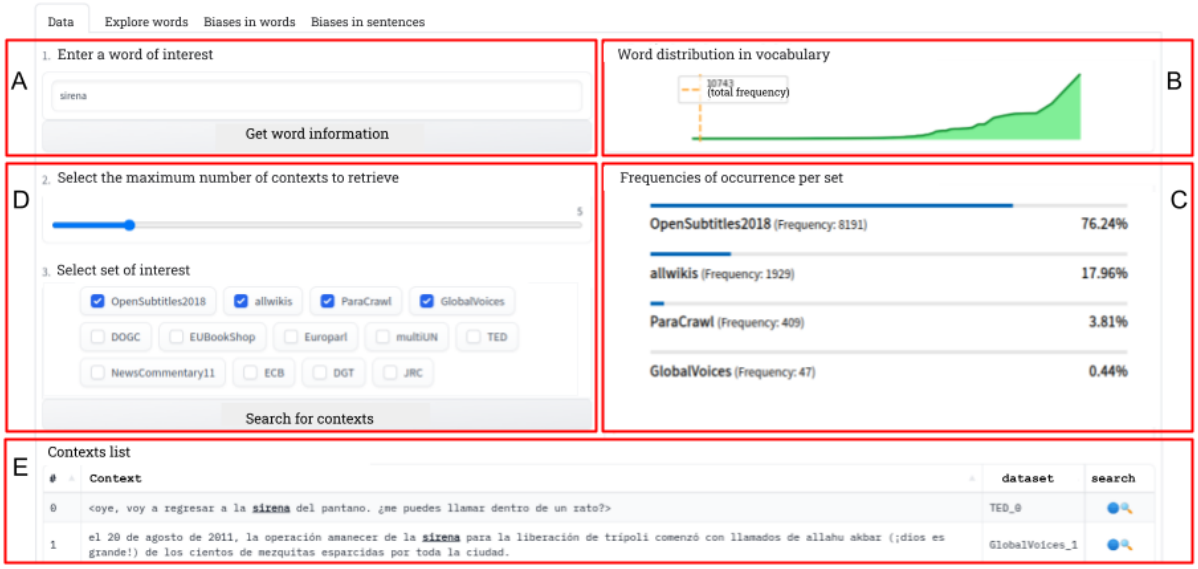


Figure 2: The Data tab of EDIA. The word of interest, selected in (A), is situated within the frequency plot of all of the words in the vocabulary in (B), and its relative frequency in different subcorpora is shown in (C). The user can retrieve contexts of occurrence of the word of interest in (D).

### 4.1 Exploration of the learning data

In hands-on experiences with discrimination experts, it was found that it was a huge priority for them to identify and study the origin of the data in detail. Indeed,

As can be seen in Figure 2, EDIA allows to explore the frequency of appearance of a word in the corpus used to train embeddings, as well as to access contexts of occurrence of those words. This allows for a more situated analysis of the word, to detect ambiguities and possible inadequate representations due to low frequencies.

### 4.2 Exploring the distribution of words in an embedding

This functionality, displayed in Figure 3, enables the visualization of a list of words of interest in a 2-dimensional space. This space is a more intuitive rendering of the original embedding space, obtained using PCA projection.

This visualization allows to assess the closeness (similarity) of the representations of different words, obtained from their contexts of occurrence in the training data.

Note that this assessment does not require any understanding of the methods used to obtain this visualization, such as vector space, cosine similarity, Principal Component Analysis or even embbeding. Without resorting to those concepts, users can obtain an intuitive notion of the potential behavior of words in applications using that embedding. Indeed, when working with users, we found that they could obtain very valuable insights from this visualization, which impacted in a more powerful usage of the functionality of bias in words.

We include a functionality to retrieve words that are similar to the words of interest. This is useful to detect unsuspected senses associated to a given token, and also to enlarge an initial word list.

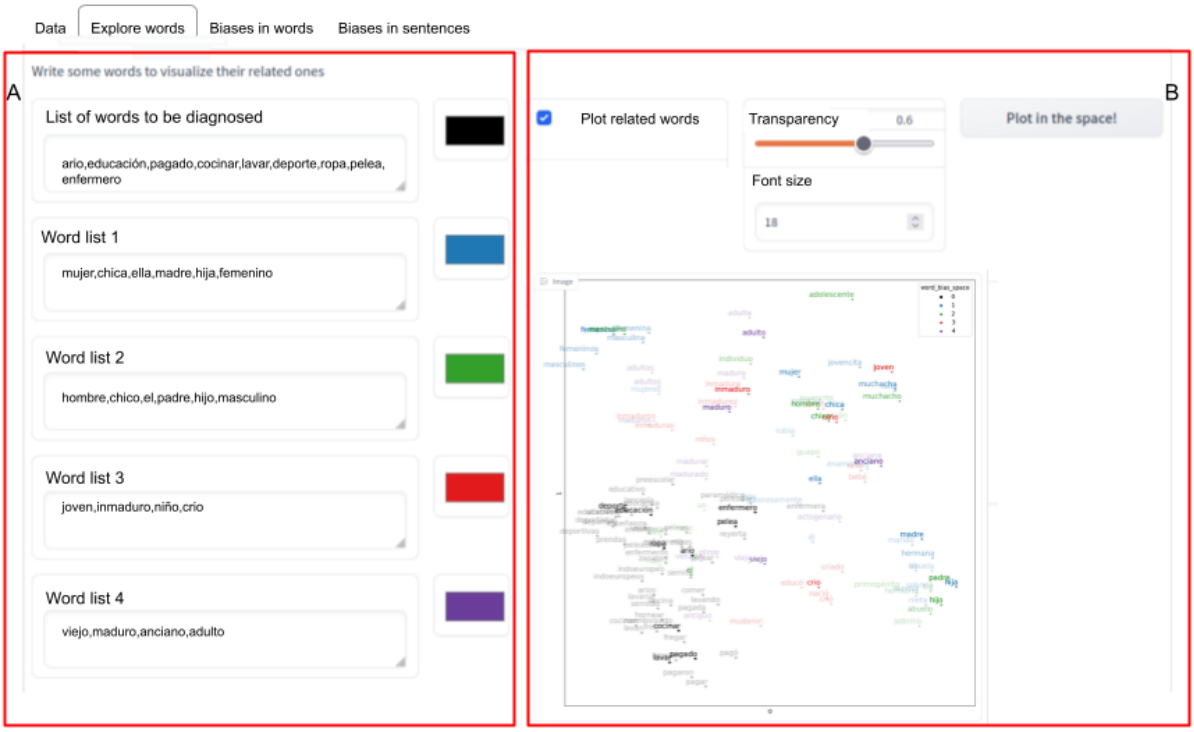


Figure 3: The Explore Words tab of EDIA. In (B), the lists of words of interest given in (A) are situated in a 2-dimensional projection of the original embedding space, obtained using PCA. Different colors are used to distinguish different lists of words. The interface also provides words that are close in the space, as suggestions.

### 4.3 Systematization of bias in words

The graphical interface to systematize the study of bias in words can be seen in Figure 4 for the case of two-bias space systematization, with a detail of the single-bias systematization shown in Figure 1.

Our core methodology to assess biases in word embeddings is iterative, relying on the feedback that the discrimination expert obtains from seeing how different words get represented in the embedding, and the adequacy of different word lists, or modifications on those word lists, to characterize the bias of interest.

The methodology is as follows:

1. Defining a **bias space**, usually binary, by defining pairs of opposed extremes, as in *male – female*, *young – old* or *high – low*. Each of the extremes of the bias space is characterized by a list of words. This list of words, shown in (A) in Figure 4 and at the top of the diagrams in Figure 1, characterizes each of the extremes of the bias, and thus the bias space. If further refinement is needed, an additional **bias space** can be defined, that can be then combined with the first one in a space with four extremes, as shown in Figure 4.

2. Assessing the behaviour of **words of interest in this bias space**, finding how close they are to each of the extremes of the bias space. Closeness is calculated with the metric proposed by Bolukbasi et al. (2016), but can be substituted by another similarity metric in the deployment of the tool. This assessment shows whether a given word is more strongly associated to any of the two extremes of bias, and how strong that association is. In Figure 1 it can be seen that the word "*nurse*" is more strongly associated to the "*female*" extreme of the bias space, while the word "*leader*" is more strongly associated with the "*male*" extreme. Such assessment allows experts to state whether a given model is biased, in this case, they would state that it is biased with respect to professions as related to gender.

3. After seeing how words of interest distribute in the 2-way (Figure 1) or 4-way (Figure 4) bias space, and looking for an adequate representation of their bias of interest, experts may decide to:

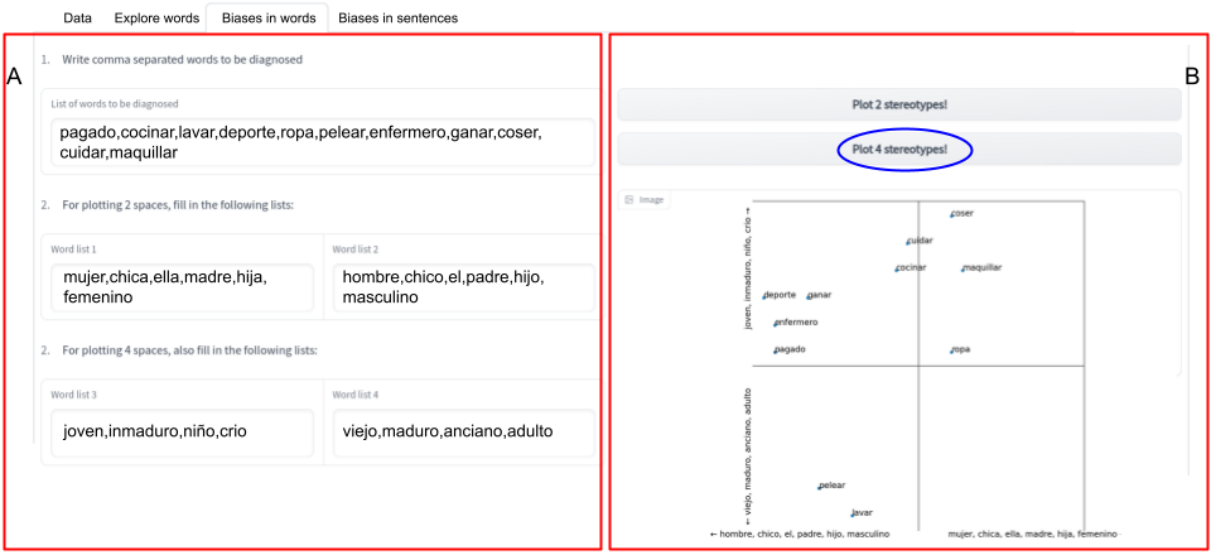


Figure 4: The "Biases in Words" tab in EDIA. The words in (A.2) shape the bias space in (B), in this case, with four extremes: one with words related to feminine, another for masculine, for old and for young. The words listed in (A.1) are positioned in (B) with closeness relative to their cosine similarity to the words in each extreme.

(a) modify some words of interest or some words in the definition of the bias extremes, possibly by resorting to exploring the distribution of words in an embedding, or exploring the training data, and going back to 2.

(b) consolidate the lists of words as a good representation of the bias of interest.

After this iterative process is finished, an assessment of bias can be produced, describing the bias in a given word embedding. This is valuable information to take informed decisions like using that embedding or looking for another one, curating the training data and retraining the embedding, or others. Moreover, the consolidated lists of words can also be used to assess that kind of bias in any other embedding.

This form of bias assessment may be useful, but in hands-on workshops discrimination experts found that it was insufficient to characterize:

- words that were highly ambiguous, like "*rico*" (rich), that can refer to economic status, flavor or part of the name of a country (Puerto *Rico*, Costa *Rica*).
- biases that were non-binary, as in gender, age, geographical origin, and many others.
- biases where one of the extremes is unmarked, as in *indigenous* - ??.

These limitations are mainly due to the fact that, in word embeddings, words are characterized in isolation. To address this limitation, the context of words needs to come into play. Thus, while the exploration of word embeddings may be useful, the exploration of language models, which is carried out via full utterances that provide context for words, is able to overcome these limitations.

### 4.4 Systematization of bias in language models

Large Language Models (LLMs) represent contextual meaning. This meaning cannot be analyzed in the analytical fashion that we have seen for word embeddings. However, LLMs can be queried in terms of preferences, that is, how probable it is that an LLM will produce a given sentence. Thus, we can assess the tendency of a given LLM to produce racist, sexist language or, in fact, language that invisibilizes or reinforces any given stereotype, as long as it can be represented in contrasting sequences of words.

Methodologies to explore bias in LLMs are proposed by Zhao et al. (2019); Nangia et al. (2020); Sedoc and Ungar (2019); Névéol et al. (2022).They are based on manually produced contrasting pairs of utterances that represent two versions of a scene, one that reinforces a stereotype and the other contrasting with the stereotype (what they call *antistereotype*). Then, the LLM is queried to assess whether it has more preference

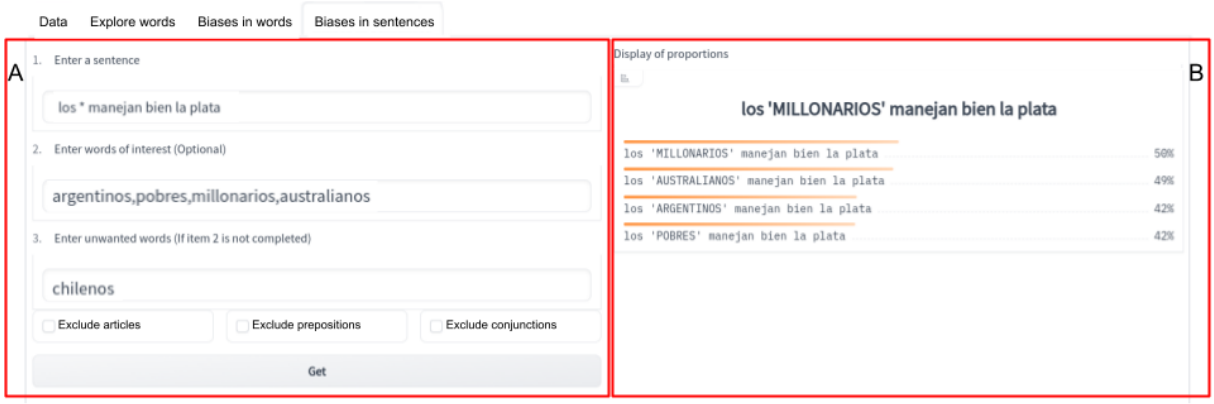


Figure 5: The "Biases in Sentences" tab in EDIA. The sentence in (A) contains a blank, represented by a "*", which is filled by each word of interest. The preferences of the model for each of the variants of the sentence are displayed in (B).

for one or the other, and how much. Such preference is calculated following the metric proposed by Nangia et al. (2020). This allows to assess how probable it is that the LLM will produce language reinforcing the stereotype, that is, how biased it is for or against the encoded stereotype.

To explore bias in sentences, EDIA provides the functionality displayed in Figure 5. The user provides a sentence with a blank (in the prototype, the blank is represented with a *). Then, the sentence is completed by filling the blank with different words, also provided by the user, that describe the different stereotypes or antistereotypes to be compared. Then, the preferences of the model to generate each of the sentences are showed. If the model shows uniform preferences, then we can state that the model has no bias with respect to the stereotypes and antistereotypes represented by the variants of the sentence. If preferences are not uniform, then some kind of bias can be assessed.

As with the exploration of word embeddings, experts can modify their lists of words and the words in the sentences, observing their probabilities in a given model, until they obtain a representation of their bias of interest that they deem adequate. The result of this iterative and interactive process is both an assessment of the model and a list of sentences that can be used to assess that same bias in other models, given that they are masked language models.

In Appendix A a detailed user story is provided, showcasing how this framework may be used.

## 5 Discussion

We have argued that bias is a complex phenomenon that needs to be addressed with specific expertise, or else risk a reductionist approach. Such approaches have been shown to produce inadequate results. To our knowledge, existing tools to address bias require technical expertise of different kinds. Such requirement will probably hinder the involvement of discrimination experts in the bias assessment problem, specially those experts belonging to minorized communities or in the Global South.

We have developed a tool, EDIA, that eliminates unnecessary technicalities. The main aim of EDIA is to facilitate that discrimination experts can build the linguistic resources (word lists and word sentences) that are the keystone of bias assessment by interacting with the relevant word embeddings and language models.

We have worked with a variety of discrimination experts in four hands-on workshops, two before the development of the graphical interface of the tool, and two after an initial prototype, involving 70 and 30 experts on diverse fields and aspects of discrimination. Experts worked in their area of expertise, and successfully modeled different biases, including ageism, fatphobia, ableism or aporafobia. They also explored stereotypes associated to the province of origin within Argentina, gender violence, the young or different psychological features. Participants were satisfied and we are planning to carry out a second phase of the project where we expect them to produce linguistic resources resulting from their systematic exploration

of those biases.

We have also carried out hands-on sessions with general public, not discrimination experts, and they have been able to use EDIA to intuitively explore biases in language models and consolidate a critical perspective on those technologies.

The work presented here is just the starting point of a much longer endeavor. Our vision is that firms and institutions integrate this kind of exploration within the development of language technologies, engaging discrimination experts as a permanent asset in their teams, well before deploying any product. We would also like the general population to carry out this kind of audits, and that this is part of a more aware, empowering technology education for all.

We are also working toward building a repository of linguistic resources that represent different biases, as characterized by different communities and in different contexts.

## 6 Limitations

For the development of our tool EDIA we designed three workshops with 50 people each in which we received feedback about its usability. We based our decisions on the feedback we received from different experts in discrimination in hands-on experiences using early prototypes of the tool. Most of the experts worked on gender discrimination and other kinds of discrimination are less represented in our workshops. For more detail on the workshops we conducted with users to assess design decisions and the overall accessibility of the methodology, see (Benotti et al., 2023).

We did not ensure that the participants in our workshops were representative of the intended population, although we did our best efforts to have people with diverse backgrounds in social and objectives. Although we did our best to have a diverse team, including social scientists, communicators, linguists and computer scientists, of diverse backgrounds, ages and geographical origins, we could not manage to integrate people with disabilities, or without university education.

Our workshops were conducted in Spanish. Our tool works for English too but the evaluation and the design was only evaluated for Spanish. We do not provide mitigation strategies in our tool, we only make bias assessment available for not experts.

Our assessment of bias in word embeddings is limited to a binary representations of bias. We allow for a more nuanced analysis of biases by combining two binary biases, characterized by four extremes (feminine vs masculine, old vs young, etc), as displayed in Figure 4. The assessment of bias in large language models, through sentences, overcomes this limitation.

## 7 Ethical Considerations

Our tool can benefit researchers from social sciences that want to study biases in word embeddings or language models. It can also be used by small companies that cannot train their own language models and that want to study the biases present in different pre trained language models when deciding which to use in their products.

The metrics we use to measure bias are known to have limitations (Badilla et al., 2021) and the benchmarks existing in the area (Blodgett et al., 2021). A potential risk of our tool is that users assume that our tool can be used to show that a model is not biased in a particular dimension without considering the limitations of the metrics and the benchmarks.

Finally, this work discusses how to involve discrimination experts in the exploration of biases in NLP and argues that this is important. This might discourage researchers in NLP working on bias analysis and mitigation to keep working in this area because they do not have access to interdisciplinary experts. In this way, we could discourage work in an area we believe is important. We think different approaches are valuable in this area and studying in more detail the metrics of the area is very important and needs deeper technical expertise. This might not require discrimination experts if reliable benchmarks are available in the area.

Participation in our workshops involved answering a pre-survey, a post-survey, and a 3-hour hands-on in-person workshop. Participants were volunteers and did not receive compensation.

EDIA does not censor the models, so words that might be censored by other tools can be explored. In one of our workshops the participants explored words associated to feminine sexuality vs words associated with masculine sexuality and found that feminine words were associated with disease while sexual masculine words were associated with health in the language model (Cañete et al., 2020).

## References


Haozhe An, Xiaojiang Liu, and Donald Zhang. 2022. Learning bias-reduced word embeddings using dictionary definitions. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1139–1152, Dublin, Ireland. Association for Computational Linguistics.

Maria Antoniak and David Mimno. 2021. Bad seeds: Evaluating lexical methods for bias measurement. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1889–1904, Online. Association for Computational Linguistics.

Pablo Badilla, Felipe Bravo-Marquez, and Jorge Pérez. 2020. WEFE: the word embeddings fairness evaluation framework. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 430–436.

Pablo Badilla, Felipe Bravo-Marquez, and Jorge Pérez. 2021. Wefe: The word embeddings fairness evaluation framework. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20.

Luciana Benotti, Laura Alonso Alemany, Hernán Maina, Lucía González, Mariela Rajngewerc, Lautaro Martínez, Jorge Sánchez, Mauro Schilman, Guido Ivetta, Alexia Halvorsen, Amanda Mata Rojo, Matías Bordone, and Beatriz Busaniche. 2023. A methodology to characterize bias and harmful stereotypes in natural language processing in latin america.

Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. Fairlearn: A toolkit for assessing and improving fairness in AI. Technical Report MSR-TR-2020-32, Microsoft.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Online. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4356–4364, Red Hook, NY, USA. Curran Associates Inc.

Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. 2019. Understanding the origins of bias in word embeddings. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 803–811. PMLR.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Cristian Cardellino. 2019. Spanish Billion Words Corpus and Embeddings.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *Workshop Practical Machine Learning for Developing Countries: learning under limited resource scenarios at International Conference on Learning Representations (ICLR)*.

Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A survey of race, racism, and anti-racism in NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1905–1925, Online. Association for Computational Linguistics.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Bhavya Ghai, Md Naimul Hoque, and Klaus Mueller. 2021. Wordbias: An interactive visual tool for discovering intersectional biases encoded in word embeddings. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, New York, NY, USA. Association for Computing Machinery.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *NAACL-HLT*.

Hila Gonen, Yova Kementchedjhieva, and Yoav Goldberg. 2019. How does grammatical gender affect noun representations in gender-marking languages? In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 463–471, Hong Kong, China. Association for Computational Linguistics.

Wei Guo and Aylin Caliskan. 2021. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, page 122–133, New York, NY, USA. Association for Computing Machinery.

Yue Guo, Yi Yang, and Ahmed Abbasi. 2022. Auto-debias: Debiasing masked language models with automated biased prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1012–1023, Dublin, Ireland. Association for Computational Linguistics.

Shlomi Hod. 2018. Responsibly: Toolkit for auditing and mitigating bias and fairness of machine learning systems. [Online; accessed <today>].

Shengyu Jia, Tao Meng, Jieyu Zhao, and Kai-Wei Chang. 2020. 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*, pages 2936–2942. Association for Computational Linguistics.

Masahiro Kaneko and Danushka Bollegala. 2021. Dictionary-based debiasing of pre-trained word embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 212–223, Online. Association for Computational Linguistics.

Sayash Kapoor and Arvind Narayanan. 2022. Leakage and the reproducibility crisis in ml-based science.

Anne Lauscher and Goran Glavaš. 2019. Are we consistently biased? multidimensional analysis of biases in distributional word vectors. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 85–91, Minneapolis, Minnesota. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Aurélie Névéol, Yoann Dupont, Julien Bezançon, and Karën Fort. 2022. French CrowS-pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8521–8531, Dublin, Ireland. Association for Computational Linguistics.

Malvina Nissim, Rik van Noord, and Rob van der Goot. 2020. Fair is better than sensational: Man is to doctor as woman is to doctor. *Computational Linguistics*, 46(2):487–497.

Flavien Prost, Nithum Thain, and Tolga Bolukbasi. 2019. Debiasing embeddings for reduced gender bias in text classification. In *Proceedings of the First Workshop on Gender Bias*

*in Natural Language Processing*, pages 69–75, Florence, Italy. Association for Computational Linguistics.

Archit Rathore, Sunipa Dev, Jeff M. Phillips, Vivek Srikumar, Yan Zheng, Chin-Chia Michael Yeh, Junpeng Wang, Wei Zhang, and Bei Wang. 2021. Verb: Visualizing and interpreting bias mitigation techniques for word representations.

João Sedoc and Lyle Ungar. 2019. The role of protected class word lists in bias identification of contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 55–61, Florence, Italy. Association for Computational Linguistics.

Ian Tenney, James Wexler, Jasmijn Bastings, Tolga Bolukbasi, Andy Coenen, Sebastian Gehrmann, Ellen Jiang, Mahima Pushkarna, Carey Radebaugh, Emily Reif, and Ann Yuan. 2020. The language interpretability tool: Extensible, interactive visualizations and analysis for NLP models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 107–118, Online. Association for Computational Linguistics.

James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson, editors. 2019. *The What-If Tool: Interactive Probing of Machine Learning Models*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. 2019. Examining gender bias in languages with grammatical gender. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5276–5284, Hong Kong, China. Association for Computational Linguistics.

## A A user story showcasing how this tool may be used

In this Section we describe a user story that presents a paradigmatic process of bias exploration and assessment.

We would like to note that this user story was originally developed to be situated in Argentina, the local context of this project. It was distilled from experiences with data scientists and experts in discrimination that are described in (Benotti et al., 2023). However, in order to make understanding easier for non-Spanish speaking readers, we adapted the case to work with English, and consequently localized the use case as if it had happened in the United States.

**The users.** Marilina is a data scientist working on a project to develop an application that helps the public administration to classify citizens' requests and route them to the most adequate department in the public administration office she works for. Tomás is a social worker within the non-discrimination office, and wants to assess the possible discriminatory behaviours of such software.

**The context.** Marilina addresses the project as a supervised text classification problem. To classify new texts from citizens, they are compared to documents that were manually classified in the past. New texts are assigned the same label as the document that is most similar. Calculating similarity is a key point in this process, and can be done in many ways: programming rules explicitly, via machine learning with manual feature engineering or by deep learning, where a key component is word embeddings. Marilina observes that the latter approach has the least classification errors on the past data she separated for evaluation (the so called test set). Moreover, deep learning seems to be the preferred solution these days, it is often presented as a breakthrough for many natural language processing tasks. So Marilina decides to pursue that option.

An important component of the deep learning approach she uses are word embeddings. Marilina decides to try a well-known word embedding, pretrained on Wikipedia content. When she integrates

it in the pipeline, there is a boost in the performance of the system: more texts are classified correctly in her test set.

**Looking for bias.** Marilina decides to look at the classification results beyond the figures of classification precision. Being a descendant of Latin American immigrants, she looks at documents related to this societal group. She finds that applications for small business grants presented by Latin American immigrants or citizens of Latin American descent are sometimes erroneously classified as immigration issues and routed to the wrong department. These errors imply a longer process to address these requests in average, and sometimes misclassified requests get lost. In some cases, this mishap makes the applicant drop the process.

**Finding systematic errors.** Intrigued by this behaviour of the automatic pipeline, she makes a more thorough research into how requests by immigrants are classified, in comparison with requests by non-immigrants. As she did for Latin American requests, she finds that documents presented by other immigrants have a higher error rate than the non immigrants requests. She suspects that other societal groups may suffer from higher error rates, but she focuses on Latin American immigrants because she has a better understanding of the idiosyncrasy of that group, and it can help her establish a basis for further inquiry. She finds some patterns in the misclassifications. In particular, she finds that some particular business, like hairdressers or bakeries, accumulate more errors than others.

**Finding the component responsible for bias.** She traces the detail of how such documents are processed by the pipeline and finds that they are considered most similar to other documents that are not related to professional activities, but to immigration. The word embedding is the pipeline component that determines similarities, so she looks into the embedding with the EDIA toolkit[3]. She defines a bias space with "*Latin American*" in one extreme and "*North American*" in the other, and checks the relative position of some professions with respect to those two extremes, as can be seen in Figure 6, on the left. This graph is generated using the button called "Find 2 stereotypes" in the tab. She finds that, as she suspected, some of the words related to the professional field are more strongly related to words related to Latin American than to words related to North American, that is, words like "*hairdresser*" and "*bakery*" are closer to Latin American. However, the words more strongly associated to North American do not correspond to her intuitions. She is at a loss as to how to proceed with this inspection beyond the anecdotal findings, and how to take action with respect to the findings. That is when she calls for help to the non-discrimination office.

**Assessing harm.** The non-discrimination office appoints Tomás for the task of assessing the discriminatory behavior of the software. Briefed by Marilina about her findings, he finds that misclassifications do involve some harm to the affected people that is typified among the discriminatory practices that the office tries to prevent. Misclassification implies that the process takes longer than for other people, because they need to be reclassified manually before they can actually be taken care of. Sometimes, they are simply dismissed by the wrong civil servant, resulting in unequal denial of benefits. In many cases, the mistake itself has a negative effect on the self-perception of the issuer, making them feel less deserving and discouraging the pursuit of the grant or even the business initiative. Tomás can look at the output of the system, but he cannot see a rationale for the system's misclassifications, he doesn't know how the automatic classification works.

**Detecting the technical barrier.** Tomás understands that there is an underlying component of the software that is impacting in the behaviour of classification. Marilina explains to him that it is a pretrained word embedding, and that a word embedding is a projection of words from a sparse space where each context of co-occurrence is one of thousands of dimensions into a dense space where there are less dimensions, obtained with a neural network. She says that each word is a vector with numbers in each of those dimensions. Tomás feels that understanding the embedding is beyond his capabilities. Then Marilina explains to him that words are represented as a summary of their contexts of occurrence in a corpus of texts, but this cannot be directly seen, but explored using similarity between words, so that more similar words are closer.

**Finding an intuitive tool for bias exploration.** She shows him some of the tools available to as-

[3] https://github.com/fvialibre/edia

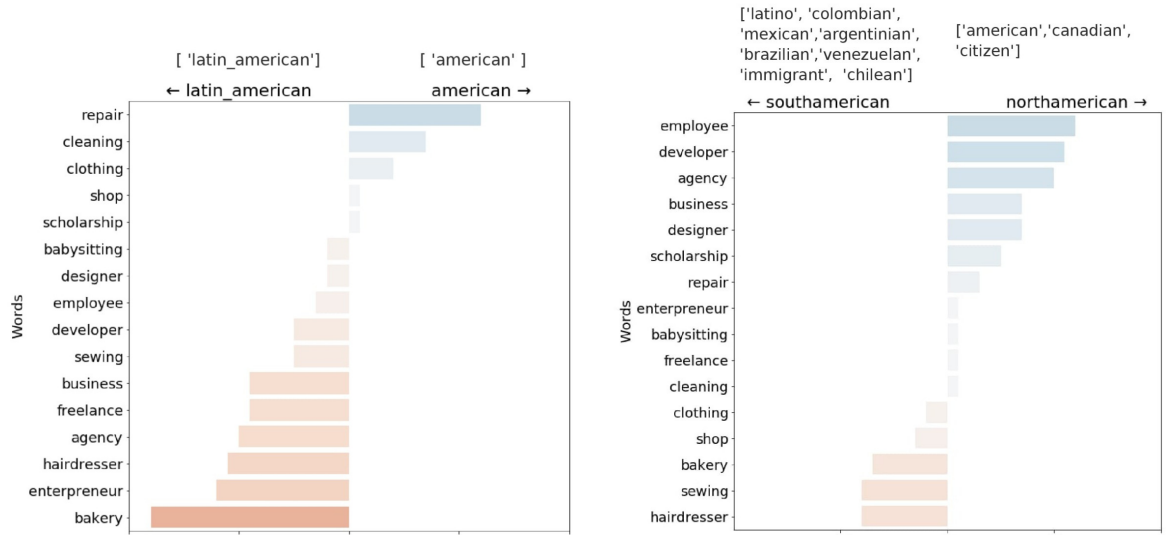


Figure 6: Different characterizations of the space of bias "*Latin American*" vs "*North American*", with different word lists created by a data scientist (left) and a social scientist (right), and the different effect to define the bias space as reflected in the position of the words of interest (column in the left).

sess bias in the EDIA demo[4], which do not require Tomás to handle any programming or seeing any code. Marilina resorts to the available introductary materials for our tool to explain bias definition and exploration easily to Tomás using the "Biases in words" tab. He quickly grasps the concepts of bias space, definition of the space by lists of words, assessment by observing how words are positioned within that space, and exploration by modifying lists of words, both defining the space and positioned in the space using the "explore words" tab with words that he know are representative for their domain. He gets more insights on the possibilities of the techniques and on possible misunderstandings by reading examples and watching the short tutorials that can be found with the tool. He then understands that word ambiguity may obscure the phenomena that one wants to study if exploring single words, that word frequency has a big impact, and that language-specific phenomena, like grammatical gender or levels of formality, need to be carefully taken into account. He uses the tab "biases in sentences" when words are highly ambiguous or when he needs to express a concept using multiword expressions such as in "Latin America". After some toying with the demo, Tomás believes this tool allows him to adequately explore biases, so Marilina deploys a local instance of the tool, which will allow Tomás to assess the embedding that she is actually using in her development, and the corpus it has been trained on.

**Explore the corpus behind the embeddings.** To begin with, Tomás wants to explore the words that are deemed similar to "*Latin American*", because he wants to see which words may be strongly associated to the concept, besides what Marilina already observed. He uses the "data tab" of EDIA, described in Section 4 to explore the data over which the embedding used by Marilina has been inferred. He finds that the embedding has been trained with texts from newspapers. Most of the news containing the word Latin American deal with catastrophes, troubles and other negative news from Latin American countries, or else portray stereotyped Latin Americans, referring to the typical customs of their countries of origin rather than to their facets as citizens in the United States. With respect to business and professions, Latin Americans tend to be depicted in accordance with the prevailing stereotypes and historic occupations of that societal group in the States, like construction workers, waiters, farm hands, etc. He concludes that this corpus, and, as a consequence, the word embedding obtained from it, contains many stereotypes about Latin Americans which are then related to the behaviour of the classification software, associating certain professional activities and demographic groups more strongly with immigration than with business. Marilina

[4] https://huggingface.co/spaces/vialibre/edia

says that possibly they will have to find another word embedding, but he wants to characterize the biases first so that he can compare to other word embeddings.

**Formalize a starting point for bias exploration.** Tomás builds lists of relevant words, with the final objective to make a report and take informed action to prevent discriminatory behavior. First, he builds the sets of words that will be representing each of the relevant extremes of the bias space. He realizes that Marilina's approach with only one word in each extreme is not quite robust, because it may be heavily influenced by properties of that single word. That is why he defines each of the extremes of the bias space with longer word lists, and experiments with different lists and how they determine the relative position of his words of interest. Words of interest are the words being positioned in the bias space, words that Tomás wants to characterize with respect to this bias because he suspects that their characterization is one of the causes for the discriminatory behavior of the classification software.

To find words to include in the word lists for the extremes, Tomás resorts to the functionality of finding the closest words in the embedding. Using "*Latin American*" as a starting point, he finds other similar words like "*latino*", and also nationalities of Latin America using the "Explore Words" tab.

He also explores the contexts of his words of interest. Doing this, he finds that "*shop*" occurs in many more contexts than he had originally imagined, many with different meanings, for example, short for Photoshop. This makes him think that this word is probably not a very good indicator of the kind of behavior in words that he is trying to characterize. He also finds that some professions that were initially interesting for him, like "*capoeira trainer*" are very infrequent and their characterization does not have a correspondence with his intuition about the meaning and use of the word, so he discards them.

Finally, he is satisfied with the definition provided by the word lists that can be seen in Figure 6, right. With that list of words, the characterization of the words of interest shows tendencies that have a correspondence with the misclassifications of the final system: applications from hairdressers, bakers, dressmakers of latino origin or descent are misclassified more often than applications for other kinds of businesses.

Even though they are assessing biases in a word embedding, that represents words in isolation, collapsing all senses of a word, Tomás believes that once they are characterizing this bias, they may best take advantage of the effort and also build a list of sentences characterizing the same bias, to be used when assessing this same bias in a language model, for example, to assess the behavior of a chatbot. To provide him with inspiration, Marilina offers Tomás a benchmark for bias exploration developed for English and French (Névéol et al., 2022) and Tomás uses that dataset partially to define his own list of sentences to explore relevant biases in this domain.

**Report biases and propose a mitigation strategy** . With this characterization of the bias, Tomás can make a detailed report of the discriminatory behavior of the classification system. From the beginning, he suspected the cultural and social reasons behind the errors, which affect more often people of Latin American descent applying for subsidies for a certain kind of business. However, his intuitive manipulation of the underlying word embedding allowed him to find words and phrases that give rise to the pattern of behavior he was observing, going beyond the cases that he has actually been able to see as misclassified by the system, and predicting other cases.

Moreover, understanding the pattern of behavior allowed him to describe properties of the underlying corpus that would be desirable in order to find another word embedding. He can propose strategies like editing the sentences containing hairdressers, designers and bakers to show a more balanced mix of nationalities and ethnicities in them. Finally, he has a list of words and sentences that can give Marilisa to measure and compare the biases with respect to these aspects in other word embeddings

## B A comparison of frameworks for bias exploration

Multiple frameworks were developed in the last years for bias analysis. Most of them require mastery of machine learning methods and programming knowledge.

WordBias (Ghai et al., 2021) is a framework that aims to analyze embeddings biases by defining lists of words. In WordBias, new variables and lists of words may be defined. This framework allows the analysis of intersectional bias. The bias

evaluation is done by a score based on cosine distance between vectors and does not allow the incorporation of other metrics. Until October 2022, this framework is only available to analyze the word2vec embedding, without having the possibility to introduce other embeddings or models.

The Visualizing of embedding Representations for deBiasing system (VERB) (Rathore et al., 2021) is an open-source graphical interface framework that aims to study word embeddings. VERB enables users to select subsets of words and to visualize potential correlations. Also, VERB is a tool that helps users gain an understanding of the inner workings of the word embedding debiasing techniques by decomposing these techniques into interpretable steps and showing how words representation change using dimensionality reduction and interactive visual exploration. The target of this framework is, mainly, researchers with an NLP background, but it also helps NLP starters as an educational tool to understand some biases mitigations techniques in word embeddings.

The What-if tool (Wexler et al., 2019) is a framework that enables the bias analysis corresponding to a diverse kind of data. Although it is not focused on text data it allows this type of input. What-if tool offers multiple kinds of analysis, visualization, and evaluation of fairness through different metrics. To use this framework researchers with technical skills will be required to access the graphic interface due to is through Jupyter/ Colab Notebooks, Google Cloud, or Tensorboard, and, also, because multiple analysis options require some machine learning knowledge (e.g, selections between AUC, L1, L2 metrics). Own models can be evaluated but since it is not text-specific, it is not clear how the evaluation of words or sentences will be. This tool allows the evaluation of fairness through different metrics.

The Language Interpretability Tool (LIT) (Tenney et al., 2020) is an open-source platform for visualization and analysis of NLP models. It was designed mainly to understand the models' predictions, to explore in which examples the model underperforms, and to investigate the consistency behavior of the models by analyzing controlled changes in data points. LIT allows users to add new datapoints on the fly, to compare two models or data points, and provides local explanations and aggregated analysis. However, this tool requires extensive NLP understanding from the user.

Badilla et al. (2020) is an open source Python library called WEFE which is similar to WordBias in that it allows for the exploration of biases different to race and gender and in different languages. One of the focuses of WEFE is the comparison of different automatic metrics for biases measurement and mitigation. As WEFE, FairLearn (Bird et al., 2020) and responsibly (Hod, 2018) are Python libraries that enable auditing and mitigating biases in machine learning systems. However, in order to use these libraries, python programming skills are needed as it doesn't provide a graphical interface.

In sum, available frameworks, even if aimed to facilitate access to existing techniques, still require some knowledge of mathematical concepts and the metrics involved. Such requirements often work as barriers for non-technical profiles.

As an alternative, we have developed EDIA, a no-code, no-statistics tool for experts to explore biases. EDIA implements metrics for bias assessment in word embeddings (Bolukbasi et al., 2016) and in language models (Nangia et al., 2020) that have well-known caveats. However, in EDIA metrics are not central, but a tool for experts to explore associations in these artifacts. They are not determinant of actions to be taken, and can be replaced by more adequate approaches, when they are available, without substantial change in the methodology of work.