Bias Mitigation in Corpora for LLMs Training Applied to Text Simplification

Victoria Muñoz-García

GPLSI research group, University of Alicante, Ctra. San Vicente s/n, 03690, San Vicente del Raspeig, Alicante, Spain
Valencian Graduate School and Research Network of Artificial Intelligence

Abstract

Large Language Models (LLMs) are trained on extensive data, and their performance and behaviour are shaped by the quality and diversity of this data. However, texts used for training these Artificial Intelligence (AI) systems often fail to reflect the wide range of human experiences and identities, which can result in reflecting and amplifying biases. Consequently, if a LLM is trained on biased texts, it can magnify biases when generating text. Regarding the continuous generation of information, it complicates the understanding of such texts, specially to users with cognitive impairment. This issue has made it necessary to resort to automated processes to make information available to all users. Therefore, Automatic Text Simplification (ATS) arises with the aim of approaching the challenge of automatically transforming an original text into a simplified and easier one. More specifically, this task will focus on the medical domain, to make medical texts accessible to all society.

Keywords

Natural language Processing, Bias mitigation, Text Simplification, Language Models, Medical texts,

1. Introduction

Based on the Responsible AI Index Report 2024, there are several dimensions that are a key part of the responsible Artificial Intelligence (AI) concept: data governance, explainability, fairness, privacy, security, safety and transparency.

When talking about creating fair and unbiased language models, fairness. Fairness refers to the development of algorithms that are equitable, avoid bias and discrimination, and take into account the diverse needs and circumstances of all stakeholders, aligning with norms of equity [1].

Machine learning methods may not only reflect biases present in our society, but also amplify them, as states Consuegra-Ayala et al. [2], following a process similar to the one illustrated in Figure 1. Therefore, it is of great importance to have quality corpora, that is to say, fair and explainable corpora, as they will have a great impact on research results.

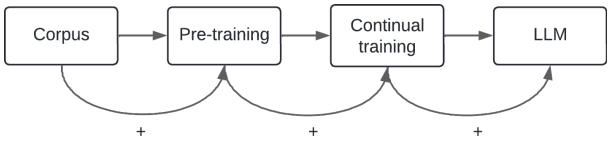

Figure 1: Bias amplification process

The daily generation of large volumes of information has meant that such biases are widespread and are used as training data for language models. In addition, the use of complex information has made it difficult for most citizens to understand relevant issues. Official communications must be accessible to all, including people with reading difficulties and/or users with cognitive impairment. Manual adaptation proves highly expensive due to the time and expertise needed to generate simplifications. Consequently, creating simplified versions of the existing volume of textual information manually

Doctoral Symposium on Natural Language Processing, 26 September 2024, Valladolid, Spain.
victoria.munoz@ua.es (V. Muñoz-García)
0009-0001-5834-9374 (V. Muñoz-García)

would be unfeasible, as in Bott and Saggion [3]. This issue has led to the need for automated procedures as to ensure that all users can access information. Therefore, Automatic Text Simplification (ATS) aims to transform automatically original texts into simplified and easier versions. Thus, this PhD thesis will focus on the Automatic Text Simplification task. Specifically, in researching ATS on the medical domain. As it was briefly introduced, data is one of the most important aspects of ATS systems. Biased data on the medical domain could have a negative impact on society. So the objective of this thesis is to research how to mitigate the bias on medical data that will be used to train ATS systems.

Bearing these considerations in mind, this doctoral symposium is intended to outline and establish the research focus and direction for this thesis, providing a clear framework and objectives that will guide the subsequent research. This academic article is divided as follows: Section 2 shows an overview on bias detection and mitigation and ATS; Section 3 describes the hypothesis and objectives of this research; Section 4 presents the methodology to be followed; and Section 5 pose some research issues open to discussion.

2. Background and Related Work

Before presenting the research proposal, this section aims to place in context the current state-of-the-art on bias mitigation and detection, as well as Automatic Text Simplification.

2.1. Bias detection and mitigation

Corpora serve as the training base for LLM development, thus it is important to use high-quality data as it significantly influence the results to be obtained. Concerns surrounding AI systems often focus on the fairness of their outcomes, as biased or unfair results can lead to significant negative social consequences [4]. In order to mitigate such biases during system training, this study emphasizes the concept of corpus fairness, which entails ensuring that a corpus accurately represents the diversity of the population.

A fair corpus is one that provides an accurate and balanced view of the language or phenomenon under study which does not promote biases.

2.1.1. Bias in LLM

Bias is defined by the Cambridge Dictionary [5] as “the action of supporting or opposing a particular person or thing in an unfair way, because of allowing personal opinions to influence your judgement”. Therefore, bias affects anyone who is discriminated against or excluded or associated with a judgement. Over 100 cognitive biases have been identified, which can be categorized into various domains such as social, behavioral, and memory-related. Cognitive bias is when human cognition consistently generates representations that are distorted compared to objective reality. Stereotyping, in particular, involves attributing specific characteristics to individuals based on their national, ethnic, or gender groups. This means certain traits are assigned to an individual simply because they belong to a particular group [6].

Detecting and mitigating bias and unfairness are complex tasks because the definition of fairness varies across different cultures. As a result, factors such as user experience, culture, social context, history, politics, law, and ethics all shape the criteria for identifying and addressing unfairness [7].

As Mehrabi et al. [8] established, data-driven AI systems and algorithms rely on the data they are trained on. Consequently, the functionality of these algorithms is closely tied to the quality of the data. When the training data contains biases, the algorithms learn and incorporate these biases into their predictions. This means that existing biases in the data can directly influence the algorithms, resulting in biased outcomes. Furthermore, algorithms can amplify and perpetuate these biases. Even when the data is unbiased, certain design choices in the algorithms can lead to biased behaviour. The biased outcomes produced by these algorithms can then influence real-world systems and user decisions, generating more biased data for future training.

According to Garrido-Muñoz et al. [9], bias in a model often stems from multiple issues during the training process, with a primary concern being the data used to train the model. If the data source under-represents a class of a protected attribute (e.g., gender or race), the model's predictions will likely favour the most represented class, intensifying underestimation for the minority class. This unequal representation can be especially problematic in sensitive decision-making systems [9].

Therefore, bias can be classified according to their source into three main different groups [8]:

- Coming from the data (data to algorithm): Measurement Bias, Omitted Variable Bias, Representation Bias, Aggregation Bias, Sampling Bias, Longitudinal Data Fallacy, and Linking Bias.
- Coming from the algorithm (algorithm to user): Algorithmic Bias, User Interaction Bias, Popularity Bias, Emergent Bias, and Evaluation Bias.
- Caused by interaction with the user (user to data): Historical Bias, Population Bias, Self-selection Bias, Social Bias, behavioural Bias, Temporal Bias, and Content Production Bias.

On account of that, many techniques have been researched and developed with the aim to detect and mitigate biases.

2.1.2. Bias detection and mitigation techniques

Based on Gallegos et al. [10], bias mitigation techniques vary based on the stage of the large language model (LLM) workflow they operate within: preprocessing, training, intra-processing, and post-processing. Pre-processing techniques for bias mitigation aim to eliminate bias and unfairness from dataset or model inputs at an early stage. In-training mitigation techniques focus on reducing bias and unfairness while the model is being trained. Intra-processing methods adjust the model's weights or decoding behaviour without requiring additional training or fine-tuning. Post-processing techniques remove bias and unfairness from outputs generated by a LLM. All these techniques and the stages they operate within are summarised in the following Figure 2. In the first column it can be seen the mitigation stage in which biases are approached and the second column shows the mechanisms that can be used. The numbers indicated in brackets correspond to the Sections in which the stages and mechanisms are detailed in the reference paper.

Mitigation Stage	Mechanism
Pre-Processing (§ 5.1)	Data Augmentation (§ 5.1.1)
	Data Filtering & Reweighting (§ 5.1.2)
	Data Generation (§ 5.1.3)
	Instruction Tuning (§ 5.1.4)
	Projection-based Mitigation (§ 5.1.5)
In-Training (§ 5.2)	Architecture Modification (§ 5.2.1)
	Loss Function Modification (§ 5.2.2)
	Selective Parameter Updating (§ 5.2.3)
	Filtering Model Parameters (§ 5.2.4)
Intra-Processing (§ 5.3)	Decoding Strategy Modification (§ 5.3.1)
	Weight Redistribution (§ 5.3.2)
	Modular Debiasing Networks (§ 5.3.3)
Post-Processing (§ 5.4)	Rewriting (§ 5.4.1)

Figure 2: Taxonomy of Techniques for Bias Mitigation in LLMs (Figure extracted from [10])

2.2. Text Simplification

Automatic Text Simplification (ATS) is a task that aims to generate simplified texts by adapting their syntax and vocabulary to ensure readability for a specific audience [3]. ATS emerges as a technology with the potential to address the challenge of automatically transforming a text into a simplified format. It is a text-to-text generation task that makes complex texts more accessible. It aims to retain the same meaning and information, but to make it easier to read and understand. This type of system replaces difficult sentences with simpler equivalents, avoiding syntactic complexity and transforming complex structures into shorter and simpler sentences [11]. Text simplification is especially important for users with cognitive disabilities, for whom it can significantly improve accessibility.

2.2.1. Text Simplification resources

ATS can be approached from different approaches: rule-based, data-driven, or hybrid. In addition, it has also been approached from some or all linguistic levels: lexical, syntactic, semantic and stylistic. The more linguistic levels and phenomena are considered in the simplification, the more accurate the simplified text will be. Furthermore, some remarks according to different aspects of the ATS are:

1. Corpora: With respect to simplified corpora, there is a paucity of training data to address the task of automatic text simplification. A key challenge for the improvement of LLMs in the simplification task lies in the limited availability of parallel data [12]. According to [13], there are 49 existing parallel corpora on simplification. Most of these corpora are aligned at sentence level and consist of a few hundred documents or sentences. In relation to language, there are 10 corpora in Spanish, being 7 of them on the general domain, 1 of them on public administration and 2 of them on the medical domain.
2. Tools: Currently, most of the tools are rule-based (12, 44.44%), while 7 are data-based (25.93%), 7 are hybrid tools (25.93%) and one, DysWebsia, is not specified (3.70%). As for the linguistic level simplified by these tools, the vast majority (23, 85.19%) perform lexical simplifications, and 11 tools simplify exclusively at this particular level. Syntactic simplification is carried out in about half of the tools analysed (14, 51.85%). As for discourse simplification, 5 tools (18.52%) address discourse-related issues [14].
3. Computational approaches: ATS have been tackled using both conventional and artificial intelligence methods, employing rule-based or machine learning systems to analyze and enhance complex texts [15]. At present, deep learning systems are used to automatically generate simplified versions of texts, through a similar process to a machine translation. Rule-based approaches rely on linguistic expertise and tools like parsers and taggers to apply predefined rules. In contrast, data-driven approaches use scientific methods and machine learning algorithms to extract simplification rules from large datasets [16].
4. Evaluation: Currently text simplification systems are commonly assessed using automatic metrics or human ratings. As stated in [17], automatic metrics like BLEU and SARI are widely used because they do not depend on the language, as BLEU focuses on n-gram overlap and SARI on measuring the changes in words (additions, deletions, and kept words). However, some metrics, such as SAMSA or readability metrics such as Flesch-Kincaid Grade Level are language-specific, initially designed for English but adapted for other languages. Additional metrics used for evaluating ATS systems include TER, ROUGE, C-Score, and E-Score. To facilitate metric calculation and comparison, the EASSE package was developed, encompassing BLEU, SARI, and Flesch-Kincaid Grade Level metrics. On the other hand, human evaluation criteria typically assess grammar and fluency, meaning preservation, and simplicity using Likert scales or similar methods. However, it is necessary to keep in mind that both automatic metrics and human judgments have limitations.

The aforementioned sections highlight two open research areas: bias detection and mitigation, and automatic text simplification. Bias detection and mitigation involves developing better detection methods, evaluation metrics, and mitigation strategies, as LLMs are still biased. On the field of automatic

text simplification, there is scarcity of data which limits the improvement of LLMs for this task. Most tools are rule-based and mainly focus on lexical simplification. Furthermore, accurate automatic metrics are needed to evaluate simplification systems effectively. These research areas contribute to the development of AI technologies that are both fair and accessible, promoting inclusiveness and ethical standards in the field.

3. Research Description

LLMs are based on biased data, more specificlly in the case of our research, on biased textual corpora. These models amplify those biases, which has a direct impact on society. In our research, we will apply LLMs to the medical domain and focus on the task of text simplification, which could make medical information accessible to a broader audience. With these motivations extracted from the state-of-the-art, the following objectives and sub-objectives were defined.

3.1. Objectives

The main objective of this doctoral thesis is to investigate methods on the detection and mitigation of biases in corpora for Large Language Models (LLMs) training applied to Automatic Text Simplification (ATS) in the medical domain.

3.2. Specific objectives

On the basis of this objective, different sub-objectives have been established in order to provide a detailed workflow:

1. Review and study of the state-of-the-art on bias detection and mitigation methods for the creation of fair and quality corpora.
2. Review and study of the state-of-the-art on the automatic simplification of texts in general and specifically in the medical domain.
3. Creation of fair and quality corpora in general and for the task of ATS.
4. Construction of fair language models for ATS in the medical domain.
5. Evaluation of constructed corpora and LLM.
6. Contribution to the creation of resources in Spanish: language resources (guides and corpora) and LMs.
7. Contribution to the Sustainable Development Goals (SDGs): 3 (health and well-being), 5 (gender equality) and 10 (reduction of inequalities).

The aim of this thesis is to create a LLM with a fair and unbiased corpus. This approach not only contributes to the reduction of inequalities (such as gender), but also reinforces the continued commitment to the SDGs in research. As focusing on mitigating gender bias through Natural Language Processing, this research will take into account SDG 10, which aims to reduce inequalities, but also contributes to SDG 5 and SDG 3, which promote gender equality and health and well-being.

The main impact expected is to contribute to the detection and mitigation of biases in order to avoid their amplification and dissemination. In addition, we intend to improve the task of automatic text simplification based on the state-of-the-art methods in order to facilitate the reading and understanding of medical texts focused on specific groups. Therefore, facilitating access to information in the medical field and contributing to the creation of essential Spanish-language resources for technological progress and development are expected outcomes.

4. Proposed Methodology

The methodology proposed for this doctoral thesis is based on a comprehensive state-of-the-art study on bias detection and mitigation and, more specifically, on the task of automatic text simplification in

the medical field. An analysis of bias detection and mitigation methods is being carried out to ensure the quality of the corpora for LLM training. Finally, these corpora and models will be evaluated.

The methodology of this doctoral thesis proposal comprises several key steps. First, it involves a comprehensive study and review of the state of the art on bias detection and mitigation, and ATS. Following this, methods for detecting and mitigating biases in LLMs will be defined and evaluated. Subsequently, a comprehensive simplification guide will be developed to provide clear instructions for simplifying text in a . This guide will serve as a foundational tool for annotators who will be tasked with annotating a simplified corpus. To ensure the reliability and accuracy of the guideline, inter-annotator agreement checks will be conducted. This involves at least three linguist experts independently annotating the same set of texts, after which their results will be compared. High agreement among annotators will indicate that the guidelines are clear and effective, while differences will be analysed to refine and improve the guidelines further. This rigorous validation process is crucial to ensure that the simplification guide can be reliably used in further simplification processes. Bias detection and mitigation methods will then be applied to create an aligned original version-simplified version corpus, which will undergo a quality assessment. This methodology will be further applied to texts in the medical domain, and finally, a language model will be trained and evaluated. This workflow can be seen at a glance in Figure 3:

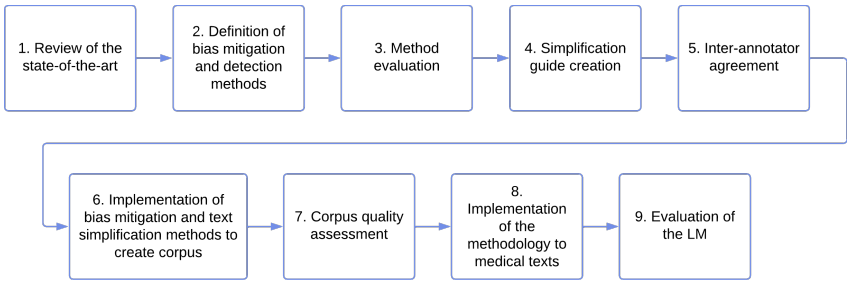

Figure 3: Proposed methodology workflow

Following this methodology, we aim to establish a general approach to balancing data and achieving fair corpora that represent real-world diversity. By creating a dataset free of biases and using appropriate metrics, we pretend to contribute to corpus quality, fairness, and explainability, thereby contributing to Responsible AI.

4.1. Related experiments and work in progress

Within the VIVES project (ref. 2022/TL22/00215334), as part of the ILENIA project, some experiments have been carried out in order to detect and mitigate biases in the first stage of a LLM lifecycle. The first step involved conducting a detailed study of the state-of-the-art. Building on this foundation, specific methods for detecting bias in text were defined: gender polarity and co-occurrences [2]. These bias detection methods were evaluated using the European Parliament's multilingual dataset, focusing on the English language. This evaluation involved applying the developed methods to the dataset and analyzing the results. The following experiments and results obtained were conducted within the work of [2].

1. Gender polarity: In order to assess gender bias in the text, the first step involved measuring the frequency of gender-related words. This analysis focused on identifying and quantifying the occurrence of words that are explicitly associated with gender. The second step involved a detailed count of masculine and feminine words, utilizing a predefined Bag of Words (BoW) approach in English. This BoW includes a curated list of gender-specific terms, allowing a systematic count and comparison of the presence of masculine and feminine words within the text. An example of the words contained in the BOW is:
    - Masculine: he, him, his, himself, man, men, he's, boy, boys

- Feminine: she, her, hers, herself, woman, women, she's, girl, girls

In the analysis of the aforementioned dataset, it was found that there are 90,542 occurrences of masculine words and 48,790 occurrences of feminine words [2]. This disparity in frequency suggests an evident imbalance in gender representation within the corpus. The higher count of masculine words compared to feminine words indicates a prevalence of gender bias towards masculine terms in English.

2. Co-occurrences: In the context of text analysis, co-occurrences refer to the frequency in which certain words appear next to each other within a given dataset. Specifically, when examining the co-occurrence of target words with gender-specific words, the analysis focuses on how often these target words appear alongside gender-specific terms [2]. Some remarkable examples were extracted in which some words such us "manager", or "ambassador" are more related to masculine words, meanwhile, "focused" or "student" appear more frequently next to feminine words, as shown in 4.

Figure 4: Some results from the European Parliament dataset (English)

This approach allowed us to identify patterns between the target words and gender-related ones, providing insights into how gender bias might be displayed in the English language.

After conducting these experiments, we aim to evaluate bias mitigation throughout various stages of a Large Language Model (LLM) lifecycle and to adapt the developed bias detection and mitigation methods specifically for Spanish.

To extend our research, we are applying these bias detection methods to Spanish texts. Currently, we are developing a comprehensive seed word list in Spanish and creating the prompts that will guide the text generation process. The generated texts will focus on topics related to human interactions and social issues, using several LLMs which are yet to be defined.

5. Research issues to discuss

This section addresses the challenges found in conducting this research that may need to be taken into consideration. Future research aims to include bias detection and mitigation methods that can be applied to the Spanish language, applying them to create quality corpora and enhance the task of ATS for medical text simplification.

Our research topics include several subjects that could be open to discussion. This research focuses on three main key topics. Here, we outline the questions that each of these topics could encompass.

1. Fairness: the first question to be addressed is the definition of fairness and quality, especially in the context of corpora and large language models (LLMs): What criteria constitute fairness and quality in these contexts, and how can these definitions be effectively applied?
2. Bias detection and mitigation: moving on to the next topic, are there comprehensive lists of features that describe various types of biases in human-related data? Can we create attribute lists for this purpose, and are there identifiable linguistic patterns that indicate and help detect biases? Are there specific patterns that need to be identified and addressed? Moving the medical domain, how can biases be minimised in health data? Is there a specific pattern to be detected?

3. Automatic Text Simplification: talking about accessibility, Who is directly affected by the difficulty of complex medical information, and which users would benefit the most from these simplification systems? Should our focus be on a specific target audience or should we aim for a broader simplification that can reach a larger number of people?

These questions form the basis for a comprehensive exploration into the intersection of fairness, bias, and text simplification in this research project.

Acknowledgments

This work has been supported by ValgrAI – Valencian Graduate School and Research Network for Artificial Intelligence and the Generalitat Valenciana https://valgrai.eu/es/ and conducted within the VIVES project (ref. 2022/TL22/00215334) https://vives.gplsi.es/.

References

[1] H.-C. A. Intelligence, Artificial intelligence index report 2024: Public data (2024).
[2] J. P. Consuegra-Ayala, Y. Gutiérrez, Y. Almeida-Cruz, M. Palomar, Automatic annotation of protected attributes to support fairness optimization, Information Sciences (2024) 120188.
[3] S. Bott, H. Saggion, Automatic simplification of spanish text for e-accessibility, in: Computers Helping People with Special Needs: 13th International Conference, ICCHP 2012, Linz, Austria, July 11-13, 2012, Proceedings, Part I 13, Springer, 2012, pp. 527–534.
[4] H. Liu, J. Dacon, W. Fan, H. Liu, Z. Liu, J. Tang, Does gender matter? towards fairness in dialogue systems, arXiv preprint arXiv:1910.10486 (2019).
[5] Cambridge university press, 2024. URL: https://dictionary.cambridge.org.
[6] I. Garrido-Muñoz, A. Montejo-Ráez, F. Martínez-Santiago, L. A. Ureña-López, A survey on bias in deep nlp, Applied Sciences 11 (2021) 3184.
[7] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, T. Gebru, Model cards for model reporting, in: Proceedings of the conference on fairness, accountability, and transparency, 2019, pp. 220–229.
[8] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM computing surveys (CSUR) 54 (2021) 1–35.
[9] I. Garrido-Muñoz, F. Martínez-Santiago, A. Montejo-Ráez, Maria and beto are sexist: evaluating gender bias in large language models for spanish, Language Resources and Evaluation (2023) 1–31.
[10] I. O. Gallegos, R. A. Rossi, J. Barrow, M. M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang, N. K. Ahmed, Bias and fairness in large language models: A survey, arXiv preprint arXiv:2309.00770 (2023).
[11] C. Scarton, Horacio saggion, automatic text simplification. synthesis lectures on human language technologies, april 2017. 137 pages, isbn: 1627058680 9781627058681, Natural Language Engineering 26 (2020) 489–492.
[12] L. Klöser, M. Beele, J.-N. Schagen, B. Kraft, German text simplification: Finetuning large language models with semi-synthetic data, 2024, p. 63 – 72. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85189859179&partnerID=40&md5=4fe3308ea5634cd4f59330e4e29df5c1.
[13] T. J. Martin, J. I. Abreu Salas, P. Moreda Pozo, A review of parallel corpora for automatic text simplification. key challenges moving forward, in: International Conference on Applications of Natural Language to Information Systems, Springer, 2023, pp. 62–78.
[14] I. Espinosa-Zaragoza, J. Abreu-Salas, E. Lloret, P. M. Pozo, M. Palomar, A review of research-based automatic text simplification tools, in: Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing, 2023, pp. 321–330.
[15] R. Alarcon, L. Moreno, P. Martínez, Easier corpus: A lexical simplification resource for people with cognitive impairments, Plos one 18 (2023) e0283622.

[16] S. S. Al-Thanyyan, A. M. Azmi, Automated text simplification: A survey, ACM COMPUTING SURVEYS 54 (2021). doi:10.1145/3442695.

[17] O. M. Cumbicus-Pineda, I. Gonzalez-Dios, A. Soroa, Linguistic capabilities for a checklist-based evaluation in automatic text simplification, volume 2944, 2021, p. 70 – 83. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85115318080&partnerID=40&md5=198bd00474b9d365db9e469c57269117.