Combining Knowledge Graphs and Node Embeddings for Gender Bias Detection and Mitigation in Greek Language

Andreas Soularidis[1,*], Dimitirs Doumanas[1] and Konstantinos Kotis[1]

[1]Intelligent System Lab, Department of Cultural Technology and Communication, University of the Aegean, Mytilene, 81100, Greece

Abstract

Gender bias excludes or promotes individuals in a plethora of fields of daily life, undermining the sense of justice and democracy. Even though a significant amount of research has already been performed on international literature, most of these studies are tailored to English texts or rely solely on stochastic methodologies such as Large Language Models (LLMs) that suffer from hallucinations and/or inaccurate responses. Motivated by our work in a recent national research project, namely FairLegal, on related research for the Greek language, in this work we propose a novel framework, namely GenderBiasRadar, to detect and mitigate gender bias in Greek documents by combining stochastic and deterministic methods. The proposed approach combines the grounding power of Knowledge Graphs (KGs), along with the utilization of Node Embeddings (NE) and Small Language Models (SLMs), to effectively detect and mitigate gender bias in Greek text. The evaluation of the proposed approach is performed to job advertisements and legal documents, while the generated responses are assessed using quantitative and qualitative metrics focusing on the cohesion and completeness of the generated KGs, the bias detection precision, and the quality (grammatical and syntactical errors) of the generated unbiased text. Despite the challenges that remain to be tackled, the preliminary experimental results are quite promising, as they demonstrate the efficacy of the proposed approach.

Keywords

knowledge graph, ontology, node embedding, gender bias, small language model

1. Introduction

The concept of bias refers to unfair and/or harmful decisions against vulnerable population groups based on race, ethnicity, gender, disabilities, etc., [1]. Particularly, gender bias is the preference or undermining of individuals towards one gender over the other [2]. The existence of (gender) bias has a detrimental impact not only on the affected individuals but on our society as a whole, undermining human rights, freedom, and democracy in general. For example, gender bias in job postings subconsciously favors or even discourages a specific gender from applying, affecting the composition in the workplace and reducing the diversity. Thus, it is crucial to find ways towards detecting and mitigating bias, promoting equal opportunities to everyone regardless of their gender.

Emerging technologies in the field of artificial intelligence (AI) and especially in Natural Language Processing (NLP), with Large Language Models (LLMs) as a primary representative, have paved new avenues in gender bias detection and mitigation. LLMs have exhibited remarkable performance in understanding and generating text. Their training in voluminous amounts of data has made them capable of understanding semantic structures and the lexical meaning for a plethora of languages [3]. This inherent capability of LLMs has led to their integration in the field of gender bias detection and mitigation [4, 5, 6]. However, despite their remarkable performance in text understanding and generation, AI systems such as LLMs, suffer from bias due to bias existing in their training data or even their training algorithm itself [7, 8]. Moreover, LLMs suffer from hallucinations and/or inconsistent responses [9], which might undermine bias detection and mitigation. Additionally, the use of LLMs raises risks concerning data privacy and protection, which in some fields, such as the legal domain,

Proceedings of the 5th International Workshop on Natural Language Processing for Knowledge Graph Construction (NLP4KGC 2026), co-located with the 25th International Semantic Web Conference (ISWC 2026), Bari, Italy, October 26, 2026

*Corresponding author.

soularidis@aegean.gr (A. Soularidis); doumanas@aegean.gr (D. Doumanas); kotis@aegean.gr (K. Kotis)

0009-0006-9769-5583 (A. Soularidis); 0009-0003-3243-5236 (D. Doumanas); 0000-0001-7838-9691 (K. Kotis)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-19

CEUR-WS.org/Vol-4335/paper8.pdf

in which the related documents contain personal/sensitive data, is of vital importance. Finally, the research and the proposed methodologies in this field are mainly restricted to a particular language, mostly in English, which cannot be applied to other languages, due to the distinct characteristics and the structure each of them has.

Motivated by the limited research in gender bias detection and mitigation in Greek text, in this paper we present GenderBiasRadar, a novel framework that fully automates this process towards generating fair - unbiased text. The proposed approach consists of two mechanisms regarding the detection and mitigation of gender bias respectively, and combines stochastic and deterministic methods by leveraging ontologies, knowledge graphs (KGs), node embeddings (NE), a Greek dictionary, and small language models (SLMs). The evaluation of the proposed framework is performed in the fields of job advertisements and legal documents (court decisions).

To ensure that sensitive and personal data that might exist in some documents, such as the legal ones, will remain private and will not be exposed to external servers and enterprises, the proposed approach is tailored to SLMs, as they allow the entire process to be performed in a local environment (e.g., local server, laptop, etc.). To this end, SLMs constitute the cornerstone of the proposed approach as they are used to automate both KG construction (text-to-KG transformation) and gender bias mitigation. On the other hand, KGs allow the representation of raw text in a structured form, as entities are represented in the form of nodes and the relations among them as edges. However, KG construction poses a significant bottleneck as it is a time-consuming task that requires extensive human effort for annotation and specialized knowledge. LLMs have proven proficiency in text understanding and generation, bringing new opportunities in the field. To this end, numerous approaches have been proposed that somehow integrate LLMs, providing automatic [10, 11] or semi-automatic (human-LLM synergy) KG construction from raw text [12, 13]. In some of them, an ontology-based approach is followed, including competency questions (CQs) generation, answering, ontology engineering, and finally KG construction [11, 12]. On the other hand, an ontology-free approach is proposed, focusing on alleviating sparsity problems by clustering the existing entities and removing or merging duplicates [10]. In contrast, a multi-modal human-LLM collaborative framework is proposed (Docs2KG), which supports NLP or on-the-fly LLM-based ontology-based generation, making the whole approach more versatile [11].

The structured format of KGs allows for efficient management and processing of information, as each node contains "targeted" and semantically structured information compared to raw (extensive) text. Moreover, KGs are used in the Named Entity Recognition (NER) approaches [14], a fundamental task in NLP that plays a crucial role in the field of legal documents, as these documents contain roles, individuals, etc., whose gender affects whether a node-sentence is biased or not. Therefore, it is evident that the creation of valid and comprehensive KGs requires the definition of a framework that will further enhance the selected models to extract the necessary entities from the given text. Ontologies play a pivotal role in the KG construction, as they define the types of entities and the relations among them in a formal manner, while they serve as blueprint or schema for the text-to-KG transformation. To this end, two ontologies were created in the presented work (one for each domain), to guide SLMs for the text-to-KG transformation.

Beyond KGs and SLMs, embeddings are used for gender bias detection. The term embeddings refers to the numerical representation of text in the form of vectors that capture the semantic and syntactic meaning, based on the context in which they appear [15]. Embeddings play a crucial role in NLP as they are the main representation of text. Thus, a plethora of embedding models have been developed and trained on huge amounts of text to capture the linguistic structure, so that semantically similar words are positioned closer in the multidimensional mathematical space. The term NE refers to the representation of KG's nodes as vectors. In this work, NEs are used to determine whether a node-sentence accommodates gender bias or not by applying metrics such as semantic (cosine) similarity. However, despite their several advantages, embeddings unintentionally include gender bias, which was transferred from the text corpora used to train these models [16]. To improve the validity and increase the accuracy of the detection mechanism of the proposed approach, the embeddings analysis results are validated - grounded by deterministic (lexicographical) analysis using a Greek dictionary.

The present work is structured around three research questions that guide our research methodology:

- RQ1: Can SLMs effectively perform the text-to-KG transformation, generating a comprehensive KG that captures the complete knowledge included in the given text document?
- RQ2: Can lexicographical analysis increase the accuracy of the gender bias detection mechanism?
- RQ3: Can SLMs effectively mitigate the gender bias present in a given text document (written in Greek) by generating the corresponding unbiased one that is also grammatically and syntactically correct?

The contribution of this work is twofold: a) it proposes a hybrid approach that combines NE with lexicographical analysis to detect gender bias in Greek texts, and b) it automates the gender bias mitigation process by introducing an SLM-based framework that detects and mitigates gender bias in Greek texts, ensuring at the same time the protection of sensitive/personal data included in them.

The remainder of this paper is structured as follows. Section 2 reviews related work on gender bias detection and mitigation strategies. Section 3 describes the proposed architecture, while Section 4 presents the research methodology and the experimental results. Finally, Section 5 discusses the findings, while Section 6 concludes the paper.

2. Related Work

Doumanas et al. [4] propose an LLM-based framework, namely DIKEOS, to detect and mitigate gender bias in Greek texts, focusing on job advertisements and legal documents. The proposed approach follows in-context learning strategies along with prompt engineering techniques applied in a wide range of language models, which vary in scale. Towards this direction, two datasets of 45 instances/examples were created, while a set of nine gender bias rules tailored to the Greek language peculiarities were defined, guiding the models to detect and mitigate gender bias. The experimental results demonstrate the supremacy of LLMs over SLMs in terms of bias detection and mitigation; however, the proposed approach is solely based on LLMs that sometimes fail to capture contextual sentence-level biases.

Khreich et al. [5] propose Genderly, a hybrid and modular data-centric system for detecting gender bias in English text by performing linguistic analysis at multiple levels. The proposed approach is not restricted to explicit gender bias, but it also identifies implicit biases within language such as sarcasm and figurative speech. The proposed framework uses traditional Machine Learning (ML) models and fine-tuned LLMS for bias detection while incorporating human-in-the-loop for validation. The experimental results demonstrate the supremacy of LLMs over traditional ML models as they excel in identifying bias embedded in semantics.

Mousavian et al. [6] leverage LLMs and prompt engineering techniques to detect and measure gender bias in passage ranking systems. Moreover, they go beyond the existing fairness metrics that mostly rely on lexical and term-based measures, by proposing a novel one, namely Class-wise Weighted Exposure (CWEx), to better quantify gender bias in ranking systems. The evaluation of the proposed approach is performed using a subset of MSMARCO collection, namely MSMGenderBias. The results indicate the influence of biases that are inherent to LLMs on fair rankings, while prompt engineering techniques such as Chain-of-thought (CoT) and few-shot prompting provide more robust results. However, the related research is oriented to ranking evaluation rather than gender bias detection and mitigation.

Ratovicius et al. [17] propose a semi-automated approach (pipeline), combining NLP, ML, and text embeddings, to assist domain experts in detecting gender bias in the legal domain. The proposed approach includes text embeddings representation and supervised binary classification, while LLMs are used to provide explanations regarding the outcomes. The evaluation was performed using annotated legal documents (judicial rulings) in the Spanish language, while the results are quite promising in terms of precision and recall. The research focuses exclusively on bias detection (in Spanish) rather than its mitigation, while the dataset used to evaluate the pipeline is limited.

Beyond the studies presented, significant research has been conducted regarding the gender bias detection and mitigation in LLMs. Particularly, in [18, 19] the authors analyze the gender bias that

is inherent to LLMs and is expressed in their responses. Similarly, in [20], the authors evaluate gender distribution produced by open-source and proprietary models, when generating personas for occupational roles including healthcare, engineering, and professional services.

Taking into account the inherent bias that exists in language models, regardless of their scale (LLMs or SLMs) and status (proprietary or open-source), the proposed approach does not exclusively relies on LLMs for gender bias detection [4, 5, 6] but it also applies stochastic and deterministic post-processing mechanisms to detect this type of bias in Greek text. Moreover, unlike [4], where the entire text is processed as a single unit, in our proposed approach the text is first transformed into a KG, or split into sentences, to facilitate its efficient processing by SLMs.

To this end, in the proposed approach, text is transformed into an ontology-driven KG, and then the corresponding NEs are generated to create the first anchor for gender bias detection using cosine similarity and two sets of words containing male and female terms. To ground the results obtained by NEs, a Greek dictionary is used, providing for each word the necessary linguistic information such as the part of speech, gender, etc. For the domain of Greek legal documents the constructed KG is used for NER implementation, capturing the entities that are present in the given text along with their roles and names.

Furthermore, concerning the bias mitigation mechanism also proposed in our approach, the selected models are informed not only about the specific word(s) inside a node/sentence that gender bias might exist, as detected by the detection mechanism, but also about how to mitigate this bias (if identified) using a set of predefined rules, as stated in [4], combined with prompt engineering techniques such as few-shot, CoT and ReAct [21]. Finally, the implementation of the proposed approach is performed in a local setup by leveraging SLMs, ensuring the privacy and safety of personal and/or sensitive data from being exposed to the Web and external servers.

3. The GenderBiasRadar Framework

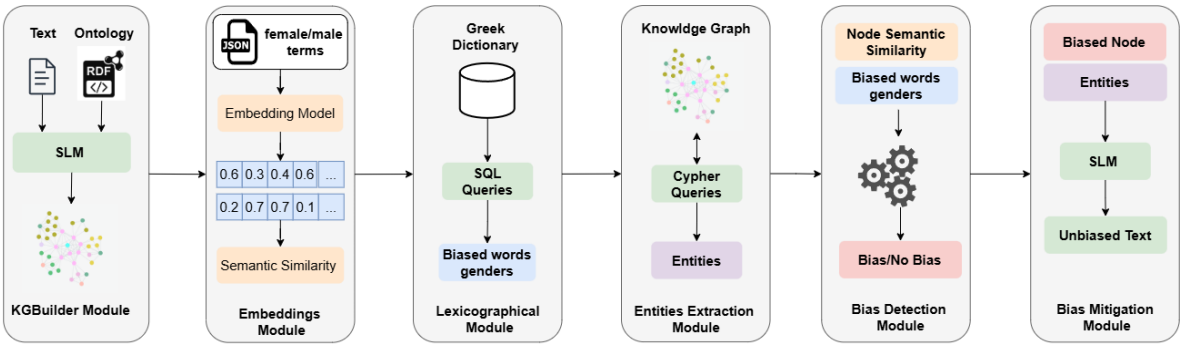


Figure 1: GenderBiasRadar high-level system architecture

The proposed framework (Fig. 1) follows a pipeline approach consisting of six distinct modules. All the related files are available at GitHub repo[1].

3.1. Knowledge Graph Construction Module

The text-to-KG transformation constitutes the cornerstone of the proposed approach, as it ensures data cleaning and consistency, splitting long sentences into nodes, etc., thereby facilitating their processing by subsequent modules. Moreover, the targeted text provided by the nodes of a KG helps SLMs effectively mitigate potential gender bias that may be present in the text. To enhance entity extraction and KG construction procedures, two ontologies were manually created (one for each domain), semantically

[1]https://github.com/AndreasSoularidis/GenderBiasRadar

describing entities and their interrelations that should be extracted from the given text. Due to the specific characteristics of each field, the created ontologies focus on accomplishing different objectives.

In more detail, in the job advertisements domain, the created ontology is based on the structure of these documents, e.g., requirements, qualifications, benefits, etc., aiming for a comprehensive and complete KG that captures all the concepts of the text. On the other hand, legal documents don't follow a strictly defined structure, while in most cases the text is extensive, containing multiple and diverse entities depending on the type of document (e.g., fraud, weapons possession, etc.). Moreover, these documents include roles along with their names (e.g., prosecutor, deputy prosecutor, defendant, etc.), which are introduced (named) once, at the beginning, and referred to throughout the text. Identifying entities (NER) such as roles, names, etc., and by extension, the gender of each entity in the text is crucial, as it largely determines whether gender bias exists or not. For example, the sentence "Ο Πρόεδρος του δικαστηρίου αποφάσισε..." (translation: "The presiding judge ruled...") contains gender bias if, earlier in the text, the role of "President" refers to a woman but doesn't contain any bias if it refers to a man. Thus, in the legal documents field, the created ontology focuses on identifying entities-roles and their related names (NER) rather than transforming the text into KG. Details of the generated ontologies are depicted in Table 1.

Table 1
Metrics of the created ontologies

Metrics	Job advertisements	Legal documents
Axioms	103	158
Classes	22	15
Object Properties	10	10
Data Properties	2	9
SubClassOf	17	5

The ontology-driven text-to-KG transformation follows an iterative approach and is powered by an SLM. Particularly, after a preprocessing step, where the text is segmented into smaller, more manageable units-sentences (chunks), multiple calls to the SLM are performed (one for each chunk). The input (prompt) of the model follows the few-shot approach and consists of a) the domain ontology, b) the sentence under transformation, c) the corresponding constructed KG, and d) text-to-KG examples (to guide the SLM). Moreover, if syntax errors are introduced, an extra validation step is performed, where the constructed graph is revised by a larger SLM, i.e., Gemma4, which in this case works as an LLM-as-a-judge, trying to evaluate, revise, and finalize the KG by fixing syntax and semantic errors that might exist. The resulting KG is finally stored at Neo4j[2].

3.2. Node Embeddings Generation Module

The objective of this module is to serve as an initial indicator for identifying potential gender bias within the nodes of the KG, using the cosine similarity metric. To this end, two sets of words are used, which include words referring to masculine and feminine genders respectively, e.g., he, him, she, her, man, woman, etc., while the embedding representation for each node (NE) is created using an appropriate sentence transformer model[3] trained on Greek text. As a next step, the cosine similarity between the centroids of both sets and the node embeddings are calculated. If the distance (overall bias score) between these two metrics are greater than a predefined threshold, then the node is considered as biased, otherwise, the node is considered as neutral. The related threshold is set to 0.02 (based on our experiments performed during the framework's development). Moreover, except for the overall bias score, the bias direction of the node is calculated. Specifically, if the calculated distance is greater than zero, the node is considered to have a bias direction toward male; otherwise, it is considered to have a

[2]https://neo4j.com/
[3]lighteternal/stsb-xlm-r-greek-transfer

bias direction toward female. To enhance the reliability of the results, a further analysis is employed at the word level, beyond the one performed at the node level, where each sentence (represented as a node) is split into distinct words where only nouns such as "κατηγορούμενος" (defendant - male), "δικηγόρος" (lawyer – male/female), etc., or adjectives e.g., "απόφοιτος" (graduate - male), "υπεύθυνη" (manager - female), etc., that typically contain gender bias are returned. Finally, the same process is followed to calculate the distance (word bias score) and the bias direction (word bias direction).

3.3. Lexicographical Validation Module

This module aims to work as a second anchor (following the embeddings module) to determine whether a node should be considered as biased or not, aiming to increase the detection accuracy and enhance the confidence of the proposed approach. To this end, an open-source Greek dictionary[4] is utilized, with more than 900.000 records and more than 518.000 distinct words. In more detail, for each word, information regarding the lemma, part of the speech (e.g., VERB, NOUN, etc.), gender (e.g., MASC, FEM, NEUTRAL), etc., is provided, allowing the gender of words to be determined in a deterministic manner.

Unlike the previous module, the evaluation is performed at word level, evaluating only the words that have been considered as biased. For each of these words, information regarding the part of speech and the gender(s) of the word are retrieved. Moreover, due to peculiarity of the Greek language, where some words, such as "δικηγόρος" (lawyer), "γιατρός" (doctor), etc., are common to both genders and the existence of an article defines the referred gender, the word (under test) and the previous one (if exists) is located into the sentence, to check whether a defined article e.g., "ο" (referred to male in Greek), "η" (referred to female in Greek), "ο/η" (referred to both genders in Greek) exists or not. The existence of the appropriate defined article before these words can determine whether inclusion is achieved or not. For example, "ο/η δικηγόρος" (the lawyer – he/she), ο/η δικαστής (the judge – he/she) ensures inclusion due to the existence of the defined article "ο/η" while "ο δικηγόρος" (the lawyer) not. Subsequently, for a word of masculine gender (as defined by the dictionary) if the defined article proceeds is "ο/η", the word is marked as non-gender biased; otherwise, it is marked as biased.

3.4. Entities Extractor Module

This module is used only in the domain of legal documents, aiming to extract the entities (individuals along with roles) from the constructed KG and exist at the text under test. For each entity, its name and role are extracted e.g., lawyer, defendant, etc., using Cypher queries that were manually created. Afterwards, the gender for each extracted entity is retrieved from the dictionary using the entity's first name as a parameter. Finally, a structure is created that for each entity includes a) the role, b) the full name, and c) the gender. This structure is used in the following modules, both in gender bias detection and mitigation in the field of legal documents.

3.5. Bias detection Module

In this module, it is finally defined whether a node-sentence accommodates gender bias or not. To this end, the analysis results obtained both from embeddings and lexicographical analysis are used. The algorithm for gender bias detection (as depicted in Algorithm 1) is common in both fields and was finalized after some experimentation using data from both fields. In more detail, for a node-sentence to be classified as biased, one of the following conditions must be met: a) the node's bias score (overall bias score) must exceed the predefined threshold, and at least one biased word must be present at the node, where the gender bias direction, as calculated in the lexicographical analysis (lexicographical bias direction) coincides with that of the sentence (overall bias direction) as calculated in the node embeddings module, and b) at least one biased word has been determined (bias score > threshold) and the gender bias direction coincides with the one calculated in the lexicographical analysis module (lexicographical bias direction) for this specific word.

[4]https://github.com/eellak/gsoc2019-greek-morpho

Algorithm 1 Bias detection algorithm

```
function DefineBias(node_analysis, threshold)
    number_of_biased_words ← length(node_analysis.biased_words)
    if node_analysis.overall_bias_score > threshold and then
        number_of_biased_words > 0 and
        overall_bias_direction matches the bias_direction of at least one word in biased_words
        return True
    end if
    if number_of_biased_words > 0 and then
        exists a word in biased_words such that bias_score > threshold
        return True
    end if
    return False
end function
```

3.6. Bias Mitigation Module

In this module an SLM is employed to mitigate the gender bias identified in the previous module, ultimately generating the corresponding bias-free text. To enhance the model's capability to effectively mitigate gender bias, intervening only in biased word(s), a combination of external knowledge integration with prompt engineering techniques, such as few-shot and CoT, is used. In more detail, among others, the prompt consists of a) the list of rules for gender bias detection and mitigation tailored to the Greek language (as described in [4]), b) a methodology (distinct steps) for gender bias detection and mitigation, c) domain-specific examples tailored to nine rules, following the ReAct approach (for legal documents), and using Retrieval Augmented Generation techniques d) a list of possible gender biased words (as defined in the previous module), and e) a list of entities along with their role, full name, and gender (for legal documents only). Finally, the model is prompted to mitigate gender bias (if exist) at the given list of possible biased words with the minimum possible intervention without adding any additional content or comments. It should be noted that this module is only triggered for those nodes-sentences where gender bias has been detected. This means that if a node-sentence is detected to be gender-biased, but it has not been detected in the previous steps of the proposed approach, the mitigation mechanism will not be triggered.

4. Evaluation of the proposed framework

The evaluation process is guided by the defined RQs (as described in Section 1). To this end, a set of job advertisements and legal documents or segments of them were used to conduct our initial evaluation. These texts were manually evaluated for possible gender bias by human experts creating the ground truth. Each experiment was conducted once and the metrics of precision, recall, and f1-score were used to measure the capability of the proposed approach to detect and mitigate gender bias in nodes-sentences. To ensure consistent results, the temperature parameter of all models used was set to 0.

In more detail, to evaluate the detection mechanism (embeddings and lexicographical module) in general and to investigate whether and to what extent the integration of deterministic methodologies (lexicographical module) affects the detection mechanism, the aforementioned metrics were calculated and compared at the end of each related module. Furthermore, a qualitative evaluation was performed regarding the text-to-KG transformation in terms of the existence of all biased instances on the constructed KG or not. To this end, two open-source SLMs i.e., Ministral 3 (14b)[5] and Gemma4 (26b)[6] were evaluated in ontology-driven text-to-KG transformation.

[5] https://ollama.com/library/ministral-3
[6] https://ollama.com/library/gemma4

Finally, to measure both the proposed gender bias mitigation module and the capacity of SLMs in successfully generating unbiased and correct text, in terms of accuracy, syntax and grammar, three open-source SLMs were evaluated i.e., Llama-Krikri (8b)[7] and OpenEuroLLM-Greek (8b), which were fine-tuned on corpora of Greek text, as well as Ministral 3 (14b)[8] that inherently supports multiple languages including the Greek one. It is noted that for the evaluation of bias mitigation module particular for the legal domain, a double approach (”loose” and ”strict”) was followed regarding the generated text. In more detail, the text generated (by SLMs) that adheres to the principle of inclusivity is considered correct even in cases inclusivity is not required (”loose” approach), because the word/role refers to a specific gender (as mentioned earlier in the text). For example, if a lawyer is introduced at the beginning of the legal text with a masculine name and the phrase “ο δικηγόρος...” (lawyer – male) is introduced afterwards and the model returns, as mitigated text, the phrase “ο/η δικηγόρος...” (lawyer – male/female), this is not considered as an error, as inclusivity is achieved. However, if the generated text contains any syntactic or grammatical errors, then it is considered as False Positive (FP). On the other hand, in the ”strict” approach, all of the aforementioned cases are considered FP, even those that are syntactically and grammatically correct.

4.1. Evaluation of text-to-KG transformation (RQ1)

As regards the text-to-KG transformation, from the experimental results it is evident that the cohesion and the completeness of the generated KGs significantly affect the bias detection mechanism (embeddings and lexicographical modules), especially in the job advertisements domain. In more detail, both models i.e., Gemma4 (26b) and Ministral 3 (12b) exhibit a comparable performance (as depicted in Table 2), and even though they manage to extract entities and their relations, defined in the given ontology from the given text, they fail to generate complete KGs that include the complete information. This deficiency is also reflected in the results, especially in the job advertisements field (Table 6). Common errors include redundant or duplicate nodes, sentence trimming of important information or the complete absence of information.

Table 2
Performance of Gemma4 (26b) and Ministral 3 (14b) in text-to-KG construction

Domain	Biased instances	Missing instances	
		Gemma4 (26b)	Ministral 3 (14b)
AD1	7	0	0
AD2	2	1	1
AD3	3	1	1
AD4	2	1	2
AD5	3	1	0

Table 3
Performance of Gemma4 (26b) and Ministral 3 (14b) in entities extraction

Document	Number of existing entities	Detected entities	
		Gemma4 (26b)	Ministral 3 (14b)
LD1	9	9	3
LD2	9	8	2
LD3	10	10	3
LD4	11	11	4
LD5	12	12	6

On the other hand, in the legal documents domain, Gemma4 (26b) exhibits a superior performance in entities extraction (Table 3), as Ministral 3 (14b) fails both to extract all the existing entities from the text,

[7]https://ollama.com/ilsp/llama-krikri-8b-instruct
[8]https://ollama.com/jobautomation/OpenEuroLLM-Greek

and to be aligned with the given ontology (mapped constructs) for the KG construction. For example, instead of annotating the name of a defendant (class Defendant) with the data property *hasValue*, that is defined in the given ontology, it uses the non-existing data property *hasName*. Thus, the related entities (expressed in triples) are not retrieved using the predefined Cypher query, undermining the mitigation mechanism.

4.2. Evaluation of gender bias detection mechanism (RQ2)

Regarding gender bias detection, the results (Table 4) demonstrate that the integration of a Greek dictionary enhances the detection's precision in both domains, increasing the total confidence of the proposed approach. In more detail, it is evident that the use of a dictionary significantly reduces the number of FP, as expected. In other words, nodes – sentences that are erroneously classified as biased by embeddings module, are correctly classified as unbiased. On the other hand, in some cases in the legal domain, sentences that are correctly classified as biased in the embeddings module, are ultimately misclassified as unbiased. This is also reflected in the metrics, as the Recall shows a significant decline. This is justified by the fact that legal documents are handled as simple text (splitted by sentence), instead of KG. Moreover, legal documents are composed of lengthy text containing multiple entities, while in some cases references to other legal documents are included, which adds "noise", such as legislative act identifiers (numbers), that affects the accuracy of the generated integrations. On the other hand, in job advertisements domain, the integration of the Lexicographical module seems to have a positive impact, as increases the total performance (f1-score) of the detection mechanism.

Table 4
Average results of gender bias detection with and without the lexicographical module

Domain	Models	Embeddings Module			Lexicographical module		
		Precision	Recall	F1-score	Precision	Recall	F1-score
Job advertisements	Gemma4 (26b)	0,25	0,61	0,35	**0,70**	0,68	**0,65**
	Ministral 3 (14b)	0,24	0,63	0,35	0,43	0,51	0,47
Legal documents	Gemma4 (26b)	0,48	**0,93**	0,61	0,62	**0,74**	0,62
	Ministral 3 (14b)	**0,55**	**0,93**	**0,65**	0,62	**0,74**	0,62

4.3. Evaluation of gender bias mitigation mechanism (RQ3)

With regards to the evaluation of the mitigation module, the experimental results (Table 5) demonstrate the superiority of Ministral 3 (14b) in both domains over smaller but fine-tuned to the Greek language models, in terms of generating unbiased text that is grammatically and syntactically correct. In more detail, in the job advertisements domain, Ministral 3 (14b) achieves better performance due to its efficacy not only to generate grammatically and syntactically correct bias-free text, but also its ability to effectively handle text that is erroneously defined as biased. Similarly, in the legal documents domain, the supremacy of Ministral 3 (14b) is based on its capability to fully exploit the given content i.e., the list of entities with their roles, names, and genders. In contrast, llama-Krikri (8b) and OpenEuroLLM-Greek (8b) show a deficiency in both handling lengthy text and exploiting the given entities, as they fail to generate unbiased text in most cases. Common errors that are also evident across all models, include their deficiency in generating unbiased text considering the given entities and/or successfully generating grammatically and syntactically correct text.

Comparing the results generated with the ones obtained by human experts, it is evident (Table 6), regarding the "strict" approach, that the proposed approach can successfully detect and mitigate half of the instances of gender bias found in these documents in both domains. In the field of job advertisements, this is mostly due to the deficiency of tested models i.e., Gemma4 (26b) and Mistral 3 (14b), to effectively extract all the necessary entities and construct a coherent and complete KG. In more detail, for the job advertisements documents, 4 out of 17 instances of gender bias are not included in the generated KG, while 3 out of 17 instances are not detected from the proposed approach. Subsequently, only 10 out of 17

Table 5
Average results of the gender bias mitigation module based on the detected bias. The average metrics of the strict approach are presented in parentheses.

Models	Job Advertisement			Legal Documents		
	Precision	Recall	F1-score	Precision	Recall	F1-score
OpenEuroLLM-Greek (8b)	0,60	0,60	0,60	0,70 (0,33)	0,45 (0,45)	0,53 (0,37)
llama-krikri (8b)	0,55	0,45	0,48	0,60 (0,35)	0,55 (0,55)	0,53 (0,40)
Ministral 3 (14b)	**0,80**	**0,65**	**0,70**	**0,81** (0,52)	**0,86** (0,86)	**0,82** (0,61)

instances of bias moved to the bias mitigation module, where only 6 out of 10 successfully transformed into unbiased and grammatically – syntactically correct text. On the other hand, in the legal domain, the detection process is solely based on embeddings and lexicographical modules, as the KG is used for NER in the mitigation module. In this case, only 20 out of 41 of biased instances are successfully detected and proceed to the bias mitigation module, where in best case i.e., using Ministral 3 (14b), 17 of them are successfully transformed into unbiased grammatically and syntactically correct text. As for the "loose" approach, the results are slightly better and highlight the limitation of the selected models (particularly the OpenEuroLLM-Greek (8b) and Llama-Krikri (8b)) to fully utilize the given list of entities to identify the corresponding roles and their genders within the given documents. Finally, the results demonstrate that in both domains the detection accuracy constitutes a critical factor in the proposed approach, while the combination of Gemma4 (26b) for text-to-KG construction with Ministral 3 (14b) for gender bias mitigation achieves the best results.

Table 6
Overall evaluation of the proposed system regarding the gender bias detection and mitigation in both domains. The average metrics of the strict approach are presented in parentheses

Models	Job Advertisement			Legal Documents		
	Precision	Recall	F1-score	Precision	Recall	F1-score
OpenEuroLLM-Greek (8b)	0,60	0,35	0,44	0,70 (0,33)	0,37 (0,37)	0,43 (0,32)
llama-krikri (8b)	0,55	0,25	0,34	0,60 (0,35)	0,47 (0,47)	0,43 (0,32)
Ministral 3 (14b)	**0,80**	**0,35**	**0,49**	**0,81** (0,52)	**0,62** (0,62)	**0,62** (0,48)

5. Discussion

The experimental results indicate the importance of each module in the proposed framework, as the degraded performance in one module significantly affects the quality and completeness of the generated results. Starting from the text-to-KG transformation, a complete, coherent and ontology-aligned KG constitutes a critical factor in the proposed approach. The results demonstrate that both models fail to integrate all the existing information into the generated KG. This weakness undermines the detection mechanisms in the job advertisements domain, where the text is transformed into KG and the mitigation mechanism in the field of legal documents, where KGs are used for NER (RQ1).

One of the key features of the proposed approach is the integration of deterministic methodologies, such as lexicographical analysis using a Greek dictionary, to address the bias that is inherent to NEs from the training phase and improve the accuracy of the detection mechanism. The results illustrate the efficacy of the proposed approach in both domains. In more detail, the integration of lexicographical analysis significantly reduces the number of FP analysis and increases the accuracy of the gender bias detection mechanism (RQ2).

Finally, regarding the ability of SLMs to effectively mitigate gender bias present in a given text (written in Greek), the experimental results demonstrate that smaller but fine-tuned to Greek text models, i.e., llama-Krikri and OpenEuroLLM-Greek, are not capable of generating unbiased text, especially in the legal documents domain, where both models fail to effectively exploit the given entities and recognize within the text. On the other hand, Ministral 3 (14b) achieves almost acceptable performance

(following the ”loose” approach), as in most cases, it effectively mitigates the detected bias, generating grammatically and syntactically correct bias-free text (RQ3). Finally, the experimental results show that the proposed approach appears, in some cases, and where the comparison is applicable, to outperform the approach proposed in [4], regarding however the use of SLMs for detecting and mitigating gender bias.

6. Conclusion

This study proposes GenderBiasRadar, a framework that automates the gender bias detection and mitigation in non-English text (Greek in our case), by combining stochastic and deterministic approaches and ensuring the protection and privacy of personal-sensitive data. The preliminary experimental results, obtained after the evaluation of the framework in job advertisements and legal documents, highlight the strengths and the weaknesses of the proposed approach. Among others, the cohesion and completeness of the generated KG from raw text significantly affects the precision of detection mechanisms, while the integration of lexicographical analysis, as a further step of validation beyond embeddings, significantly improves the precision of the proposed detection mechanism. Moreover, regarding the SLM-based mitigation mechanism, the experimental results indicate that larger models such as Ministral 3 (14b) seems to perform better compared to smaller ones, allowing the generation of grammatically and syntactically correct bias free text. Future work includes the experimentation with further language models, the refinement of text-to-KG transformation process, and experimentation with alternative sets of words and sentence models, tailored to the Greek language, for the generation of more precise embeddings from nodes-text.

7. Acknowledgments

Authors acknowledge partial support of this work by FairLegal project (MIS:YP3TA-0559990), which is funded by the National Recovery and Resilience Plan “Greece 2.0”, “SUB1.1: Clusters of Research Excellence (CREs)” with financing from the European Union—NextGenerationEU through the National Strategic Reference Framework unit of the Ministry of Education, Religious Affairs and Sports of Greece.

Declaration on Generative AI

During the preparation of this work, the author(s) used DeepL to perform Grammar and spelling check. Further, the author(s) used GitHub Copilot and Claude Sonnet 4.6 to assist in developing code implementations for the proposed framework. All AI-assisted outputs were reviewed, validated, and refined by the authors to ensure accuracy and reproducibility. The authors accept full responsibility for the content of this publication.

References

[1] T. Sun, A. Gaut, S. Tang, Y. Huang, M. ElSherief, J. Zhao, D. Mirza, E. Belding, K.-W. Chang, W. Y. Wang, Mitigating gender bias in natural language processing: Literature review, in: Proceedings of the 57th annual meeting of the association for computational linguistics, 2019, pp. 1630–1640.

[2] C. A. Moss-Racusin, J. F. Dovidio, V. L. Brescoll, M. J. Graham, J. Handelsman, Science faculty’s subtle gender biases favor male students, Proceedings of the national academy of sciences 109 (2012) 16474–16479.

[3] L. Lin, L. Wang, J. Guo, K.-F. Wong, Investigating bias in llm-based bias detection: Disparities between llms and human perception, in: Proceedings of the 31st International Conference on Computational Linguistics, 2025, pp. 10634–10649.

[4] D. Doumanas, A. Soularidis, N. Zafeiropoulos, S. Chatzistamatis, G. E. Tsekouras, A. El Saer, C. Nathanailidis, K. Kotis, Unbiasing greek: In-context learning strategies for gender bias identification and mitigation for legal documents and job ads, Information 17 (2026) 342.
[5] W. Khreich, J. Doughman, Genderly: a data-centric gender bias detection system, Complex & Intelligent Systems 11 (2025) 306.
[6] M. Mousavian, Z. Abbasiantaeb, M. Aliannejadi, F. Crestani, Towards fair rankings: Leveraging llms for gender bias detection and measurement, in: Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR), 2025, pp. 56–66.
[7] E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M.-E. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis, et al., Bias in data-driven artificial intelligence systems—an introductory survey, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 10 (2020) e1356.
[8] 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, Computational linguistics 50 (2024) 1097–1179.
[9] G. Perković, A. Drobnjak, I. Botički, Hallucinations in llms: Understanding and addressing challenges, in: 2024 47th MIPRO ICT and electronics convention (MIPRO), IEEE, 2024, pp. 2084–2088.
[10] B. Mo, K. Yu, J. Kazdan, P. Mpala, L. Yu, C. Kanatsoulis, S. Koyejo, Kggen: Extracting knowledge graphs from plain text with language models, Advances in Neural Information Processing Systems 38 (2026) 30092–30115.
[11] X. Feng, X. Wu, H. Meng, Ontology-grounded automatic knowledge graph construction by llm under wikidata schema, arXiv preprint arXiv:2412.20942 (2024).
[12] V. K. Kommineni, B. König-Ries, S. Samuel, Towards the automation of knowledge graph construction using large language models., in: NLP4KGC@ SEMANTiCS, 2024, pp. 19–34.
[13] Q. Sun, Y. Luo, W. Zhang, S. Li, J. Li, K. Niu, X. Kong, W. Liu, Docs2kg: Unified knowledge graph construction from heterogeneous documents assisted by large language models, arXiv preprint arXiv:2406.02962 (2024).
[14] H. Zhang, W. Li, D. Huang, Y. Tang, Y. Chen, P. Payne, F. Li, Kogner: a novel framework for knowledge graph distillation on biomedical named entity recognition, arXiv preprint arXiv:2503.15737 (2025).
[15] M. Siino, I. Tinnirello, M. La Cascia, Is text preprocessing still worth the time? a comparative survey on the influence of popular preprocessing methods on transformers and traditional classifiers, Information Systems 121 (2024) 102342.
[16] N. Sevim, F. Şahinuç, A. Koç, Gender bias in legal corpora and debiasing it, Natural Language Engineering 29 (2023) 449–482.
[17] C. J. Ratovicius, J. A. Diaz-Pace, A. Tommasel, Detection of gender bias in legal texts using classification and llms, in: 2024 L Latin American Computer Conference (CLEI), IEEE, 2024, pp. 1–10.
[18] Y. Wan, G. Pu, J. Sun, A. Garimella, K.-W. Chang, N. Peng, "kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters, in: Findings of the Association for Computational Linguistics: EMNLP 2023, 2023, pp. 3730–3748.
[19] V. Thakur, Unveiling gender bias in terms of profession across llms: Analyzing and addressing sociological implications, arXiv preprint arXiv:2307.09162 (2023).
[20] I. Mirza, A. A. Jafari, C. Ozcinar, G. Anbarjafari, Quantifying gender bias in large language models using information-theoretic and statistical analysis, Information 16 (2025) 358.
[21] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, Y. Cao, React: Synergizing reasoning and acting in language models, arXiv preprint arXiv:2210.03629 (2022).