Scientometrics (2025) 130:4849–4877
https://doi.org/10.1007/s11192-025-05384-1

Improving fair name-based prediction of gender in scientific communities

Maria Guariglia Migliore[1,2,3] · **Gregorio D'Agostino**[3] · **Tatiana Patriarca**[3] · **Antonio De Nicola**[3]

Received: 7 May 2024 / Accepted: 3 July 2025 / Published online: 26 August 2025

Abstract
The role of women in modern society is a central problem in several developed countries. Despite encouraging policies, women's participation in STEM fields is significantly lower than men's one. In order to develop solutions for mitigating this disparity, a deeper understanding of the underlying causes is crucial and a proper quantification of the phenomenon represents a first step to any analysis. While the problem of gender gap in scientific communities was long debated, information on authors' genders is often unavailable (see, for instance, ResearchGate and Scopus). Additionally, the lack of open-source software for automated gender prediction based on names calls for time costly human efforts. It arises the need for novel effective algorithms. Moreover, as a further challenge, desired software should guarantee gender fairness by providing the same performance for both male and female names recognition. In this paper, we propose a gender fair software to automatically predict authors' gender from their given names. The code leverages most of the existing information sources, i.e., Scopus, Semantic Scholar, and Harvard dataset. We performed an experimental application by analysing two datasets of publications, thus providing interesting insights. Finally, we evaluated the software performances in terms of accuracy, precision, recall, F1-score, and gender fairness by means of two distinct case studies. The proposed solution can enable fairer gender prediction by combining open data with carefully calibrated criteria, matching the performance of commercial tools while offering a transparent and accessible solution.

Keywords Bibliometric analysis · Prediction · Gender · Gender fairness

Introduction

Gender equality is recognized as one of the key objectives in building a sustainable society, as highlighted by the United Nations' 5th Sustainable Development Goal: "achieve gender equality and empower all women and girls" (United Nations & Development, 2015). However, attaining this goal necessitates a holistic approach to overcome societal barriers languished over centuries. This approach involves different aspects: analyzing the various levels of participation of women and men in societal activities (Abramo et al., 2021;

Extended author information available on the last page of the article

De Nicola & D'Agostino, 2021; Zhang et al., 2021; Huang et al., 2020; Choji et al., 2024) understanding the diverse opportunities available to different groups; raising awareness of gender gap issues; and taking action to address them. An essential initial step in this process is the analysis of different levels of participation, which helps to pinpoint those sectors requiring more focused intervention from decision-makers. For example, it is widely acknowledged that STEM (Science Technology Engineering and Mathematics) activities exhibit significant gender disparities (Payton & Berki, 2019), thus demanding more robust intervention measures.

This paper focuses on gender representation in scientific communities, where analyzing publication records is a key method for assessing participation, collaboration, and visibility. Therefore, a crucial step for analyzing gender gaps is accessing datasets containing researchers' personal information, including their gender. Unfortunately, analysts often face incomplete data, and they need to collect information scattered across different repositories such as Scopus[1] or ResearchGate.[2] Additionally, even when a researcher's first name is available, determining their gender can be challenging due to cultural and linguistic variations in naming conventions across countries and the uncertainty surrounding a scientist's nationality, especially if they have moved to another country (Deville et al., 2014). For example, the name Andrea is masculine in Italy but feminine in Germany. While tools and resources like *Gender API*[3] exist and help with gender prediction from names, most of them are commercial products, leaving users in the dark about how they work. Moreover, information on their performance with different genders (male or female names) are not provided. In this context, we define *gender fairness* in name-based prediction as the extent to which both males and females are predicted equally. Unequal performance based on gender can exacerbate bias, thereby worsening the proper detection of gender gaps. While commercial tools such as Gender API and NamSor rely on large proprietary databases and may achieve high performance, our open-source solution offers comparable accuracy in several settings and provides explicit mechanisms to manage uncertainty and fairness.

In this paper, the following research questions are addressed:

1. What methods can be employed to perform automated gender prediction from author names using open and reproducible data sources?
2. How can the fairness of gender prediction be achieved and assessed?

In this respect, we present a method and a tool designed for fair name-based prediction of author gender. The proposed approach starts relying on the DOI (Digital Object Identifier) of scientific papers that facilitates the integration of data from various sources, including Scopus, HARVARD World Gender Name Dictionary (Raffo 2021), and Semantic Scholar.[4] Then, it addresses the inherent incompleteness and uncertainty of available information and ensures gender fairness through an innovative method that utilizes precision, recall, and F1-score measurements of the software. We conducted experiments using a dataset collected during the activities of the European project gEneSys (De Nicola et al., 2024) and another dataset comprising papers presented at the International Conference on Critical Information Infrastructures Security from 2006 to 2022.

[1] Scopus web site: https://www.scopus.com
[2] ResearchGate web site: https://www.researchgate.net
[3] *Gender API* web site: https://gender-api.com
[4] Semantic Scholar web site: https://www.semanticscholar.org

According to Gautam et al. (2024), name-based gender prediction raises social and ethical concerns, as names often carry cultural associations with gender, religion, race, and ethnicity that can lead to oversimplified or incorrect assumptions. In a globalized world, naming conventions are diverse (Alford, 1987; Hough, 2016; Meganathan, 2009), non-binary (Knowles et al., 2016), and shifting over time, making it increasingly inappropriate to rely solely on names to infer a person's identity, especially as Natural Language Processing (NLP) tools often introduce errors and biases in gender categorization, which need to be taken into account. Furthermore, while gender is generally recognized as a social construct used to classify individuals as men, women, or other identities (LGBTQIA Resource Center 2024), data limitations necessitate a narrower definition. Consequently, we are constrained to adopt a binary distinction of genders. Therefore, we will utilize the terms "men/women" and "males/females", as they are commonly interchanged in current literature (Eagly et al., 2019). When gender cannot be determined, we employ the term "undetermined". These terms reflect perceived gender, inferred either through name-based prediction or manual assessment of publicly available profile information. They do not imply self-identified gender, biological sex, or any ideological or philosophical stance.

Throughout this manuscript, particularly in the figures, green is consistently associated with females, while red is associated with males and black to the undetermined cases.

The rest of the paper is organized as follows. The related work is presented next. This is followed by an introduction to the overall workflow for fair name-based gender prediction, and a description of the software architecture for gender prediction. The subsequent section covers the use cases, including the experimental research objectives, data collection process, human validation procedures, and achieved results. The following part discusses the findings and potential threats to validity. Finally, the paper concludes with some considerations and possible directions for future research.

Related work

Numerous studies have explored the reasons behind the persistent gender disparity in participation across various scientific disciplines (Santamaría and Mihaljević 2021; Sebo 2021; Gomide J, 2017; Bonham and Stefan 2017). Notably, these studies reveal a lower representation of women compared to men in STEM fields. A crucial prerequisite for conducting such research is the ability to predict accurately the gender of scientists. This section delves into the available tools for name-based gender prediction and highlights some of the most significant studies that have utilized these methods.

There are several tools available to predict gender from names. Some of the most popular ones are described below:

Gender guesser[5] is a commercial tool enhanced by artificial intelligence, which uses the morphology of names to infer the gender of a given first name. For increased precision, it also accepts an optional last name as input. According to the corporate website of *Gender guesser*, *Namsor*[6] (Bèrubè et al., 2020) provides the name-checking capabilities used by *Gender Guesser* and predicts gender based on a full name.

[5] *Gender guesser* web site: https://gender-guesser.com

[6] *Namsor* web site: https://namsor.app

Gender API is a commercial tool for retrieving gender information. It determines gender with an associated accuracy score, based on either a first name alone or in combination with a country. This system uses a database of first names

linked to gender based on government records, and provides a gender probability derived from the frequency with which a name appears in the dataset.

The *Genderize.io API*[7] is another commercial tool that enables users to assess the statistical likelihood that a name belongs to either a male or female individual (Sebo, 2021).

Wiki-Gendersort[8] is a completely free, Wikipedia-based algorithm for determining gender from a first name (Bèrubè et al.,). The algorithm first identifies and cleans content from Wikipedia pages associated with the name, and then

counts gendered keywords, such as *she*, *he*, *his*, and *her*. For instance, if the number of masculine keyword occurrences (e.g., *he* and *his*), is three times or more than the number of feminine ones, the algorithm assigns a masculine gender. The classifications include M (male), F (female), UNI (unisex), UNK (unknown), and INI (initials). Wiki-Gendersort relies on content from Wikipedia, which is continuously updated. While this means that the output of the algorithm may evolve over time, this characteristic aligns with the behavior of most commercial systems that rely on live or periodically refreshed data sources. Additionally, although Wiki-Gendersort only considers English-language pages, which may introduce some bias against international names, it benefits from Wikipedia's growing reputation as a reliable and structured resource, particularly when compared to other name inference methods that draw on less curated data, such as social media. The performance of *Wiki-Gendersort* has been evaluated against four different databases, namely, *gender-c*, *gender-checker*, *NamSor* dataset, and *U.S. Census Data*. *gender-c* is a free database that contains 46,599 names. *gender-checker* gathers 102,240 names and is based on UK Census data, UN Census data, and other online sources. *U.S. Census Data* contains names that represent approximately 90% of U.S. population.

Sebo (2021) performed a study with the goal of making a comparison performance between gender detection tools. Four tools were considered that meet 3 criteria: accept file format, to be partially free, and usable without advanced computer. Four database of physicians affiliated with the University Hospital of Geneva has been used. The total number of physicians is 6264, and the national's origin are Swiss, Germany, Italy, France, Austria. Hence, this study does not deal with Asian names, which are the most difficult to analyze. In these databases there are not information about geographic provenance. Considering that cultural context can influence the inference, it was used a tool named *nationalize.io*[9] to predict physician's nationality considering their first name. Gender detection tools were evaluated based on some metrics, which consider correct classifications, misclassifications (i.e., wrong gender assigned to a name), and nonclassifications (i.e., missed classifications). According to this study, the best tools are *GENDER API* and *NamSor*.

Santamaría and Mihaljević (2021) examinated five gender detection tools, respectively, *Gender API*, *NamSor*, *Genderize.io*, *gender-guesser*, and *NameAPI*[10]. The experimentation was done with a database including 7076 manually labelled names. These encompass also Asian names and, for this reason, there was a high risk of errors in gender assignment.

[7] *Genderize API* web site: https://genderize.io

[8] Wiki-Gendersort code and database: https://github.com/nicolasberube/Wiki-Gendersort

[9] *nationalize.io* https://nationalize.io

[10] *Name API* web site: https://www.nameapi.org

Gender Api and *NameSor* turned out to be the most precise tools. The study reveals also that *Gender API* and *Genderize.io* are the easiest to use.

More recently, Van Buskirk et al. (2023) proposed an open source method inspired by Cultural Consensus Theory (CCT) to predict gender from the names based on the Bayes Theorem and 36 publicly available sources of names and their gendered associations. Furthermore, they constructed a taxonomy to organize names based on how much data is available, how strongly the name is gendered, and whether the characteristics of the name vary over time or between countries. They presented also a comparison revealing that its performance is comparable to that of existing commercial tools (i.e., *Gender API*, *NameSor*, *OnoGraph*[11], and *Genderize.io*).

Sánchez-Jiménez et al. (2024) used CCT to assess gender disparities in academia. The principles of CCT are implemented in a Python package "nomquamgender" (NQG).

Among the tools discussed, our proposal stands out, based on our best knowledge, given the limited transparency of commercial tools, as the only one addressing the issue of fairness in gender determination to mitigate potential biases. Achieving high recall and reliable results in this context requires high-quality input data. One key challenge, common to all tools, is the occurrence of names represented only by initials and surname, which can hinder accurate gender inference (Huang et al., 2020). For instance, Larivière et al. (2013) reported that 31% of author names in the Web of Science dataset consist solely of initials, a problem that also persists in Scopus. While this issue relates to the broader task of author name disambiguation, which aims to reconstruct or enrich incomplete author metadata, it is conceptually separate from name-based gender prediction. To address this limitation in our pipeline, we incorporate a preprocessing step that queries Semantic Scholar using author identifiers to retrieve full first names prior to applying gender prediction algorithms. Additionally, the tools mentioned above are primarily commercial, with the exception of Wiki-GenderSort, the method proposed by Van Buskirk et al., and the Python package "nomquamgender". Consequently, there is limited information available regarding their operational mechanisms, result quality, and the metrics employed for evaluation. In contrast, Wiki-GenderSort is entirely free and utilizes Wikipedia pages to identify occurrences that may indicate gender, albeit most of these pages are in English. This could lead to bias against non-English first names. The selection of a tool may hinge on the dataset of names intended for experimentation. Some datasets comprise predominantly Western names, while others feature predominantly Asian names. Therefore, evaluating each tool individually is crucial since they may yield varying performances depending on the nationality of the names being tested.

Workflow for fair name-based prediction

The method devised for fair name-based prediction comprises four steps, ranging from dataset selection to the assignment of gender to a researcher. Each step of the workflow is outlined at an abstract level, while also suggesting the potential resources that can be utilized to accomplish it. They are as it follows:

1. *Dataset selection*.

[11] *https://forebears.io/onograph/* web site: https://forebears.io/onograph/

2. *Retrieval of researchers' information*.
3. *Assessment of gender likelihood*.
4. *Fair gender assignment*.

Dataset selection involves choosing a collection of papers to identify a group of researchers. The selection process depends on the specific objectives of the analysis. For instance, these objectives may include gender analysis of authors participating in a conference (such as the International Conference on Information Systems - ICIS), contributing to a scientific journal (for example, Scientometrics), or, broadly, addressing a scientific topic such as hydrogen energy, digital twins, or climate change. Papers can be chosen from online databases that aggregate references and citations from academic journals, conference proceedings, books, and other documents. Examples of such databases include Elsevier Scopus and Clarivate Web of Science. Among the various functionalities provided by these platforms, analysts can choose one or more keywords and conduct searches. Most importantly, the DOI (Digital Object Identifier) of the paper must be retrieved to identify it uniquely.

Retrieval of researchers' information deals with gathering as much information as possible to predict their gender. In some cases, when a large number of papers are retrieved from Scopus, first names can be dotted. This hinders name-based prediction. Therefore, other information sources, such as Semantic Scholar, must be consulted to obtain details such as the first name, the aliases, and the affiliations. While Scopus provides high-quality bibliometric data and robust author name disambiguation (AND) through proprietary algorithms, programmatic access to full, disambiguated author names via Scopus is restricted by licensing and API limitations. In our workflow, Scopus was primarily used to retrieve DOIs and basic publication metadata. Due to the prevalence of dotted or abbreviated first names in Scopus records, we adopted Semantic Scholar as an open-access enrichment layer, leveraging its API to obtain expanded author metadata, including full names and aliases, in a reproducible and automated manner. This decision prioritized transparency, scalability, and reproducibility. We acknowledge that Scopus's AND system is highly accurate, and future work could explore hybrid approaches that combine open-access and proprietary resources to maximize both data quality and accessibility.

Assessment of gender likelihood regards addressing the inherent uncertainty of the name-based prediction activity. Furthermore, while some online resources match names with genders, they typically require knowledge of the person's nationality. One of the most authoritative resources to this purpose is the HARVARD World Gender Name Dictionary. However, when dealing with data related to thousands or more scientists, determining their affiliations or nationality can be challenging. Consequently, gender can only be estimated probabilistically. For each researcher, we computed the likelihood that they are male, female, or undetermined by using the software and the algorithms described in Sect. 4 and included in the Appendix.

Fair gender assignment involves assigning a gender to a person based on the gender probabilities estimated in the previous step, without introducing further bias to the dataset, stemming from variations in the recognition of males and females. To achieve this goal, we calculated *precision*, *recall*, and F1-score for each gender, based on a sufficiently large number of manually verified researchers, ensuring these metrics remain unchanged as the number of verified researchers increases. Subsequently, we determined the threshold probability for assigning the gender of a researcher, aiming for a ratio of these metrics approaching 1. This ensures that the gender assignment process remains fair.

We emphasize that our proposed workflow is designed to be modular and adaptable to a range of research contexts. While we use Scopus and Semantic Scholar as illustrative examples for data retrieval, these sources are not required components of the pipeline. Users can substitute other bibliographic databases or institutional metadata sources as appropriate for their study. Similarly, the pipeline supports flexible integration with different name-based gender prediction algorithms. The final gender assignment step involves a fairness-aware thresholding procedure, which currently requires semi-automated calibration tailored to the specific dataset and user-defined fairness criteria. We have clarified these aspects to highlight the pipeline's flexibility and to address concerns about dependence on specific data sources or fully automated procedures.

Architecture of the gender prediction software

Overall architecture

From a software standpoint, the name-based prediction process operates as illustrated in Fig. 1 and elaborated below. Bibliometric data pertaining to a sample of papers is retrieved from a scientific online database, Scopus in our case, and, then, imported (1) into the *Gender Prediction Manager*. The Scopus database was created in 2004 by the publishing house Elsevier and includes more that 94 million records related to publications, their authors, and citations. Here, for each paper, the DOI is utilized to query the Semantic Scholar search engine for scientific literature (2) and gather information on the authors (3), including their full names and aliases. An alias is an alternative representation of the full name of the author as it appears in the articles. Aliases play a crucial role because, in some instances, parts of the name may be abbreviated or incomplete. Instead, aliases provide several versions of the name, from which the module selects the appropriate part for prediction. Subsequently, the *Gender Prediction Manager* extracts the name from the full name. The previously acquired HARVARD World Gender Name Dictionary (WGND 2.0) is employed (4) to correlate the name with a gender probability. The WGND 2.0 is a compilation of names from various countries, organized based on gender. During the execution of this workflow, the sample of papers may undergo further pruning as Semantic Scholar might fail to provide the full names of the authors for each paper. For each author, the software is

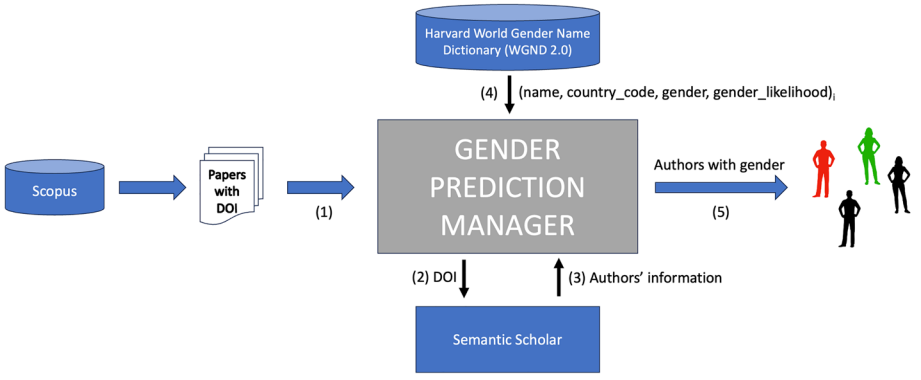

Fig. 1 Architecture of the gender prediction software

capable of associating three gender likelihoods (5): the likelihood that the author is male, the likelihood that they are female, and the likelihood that their gender is undetermined.

Semantic Scholar was founded in 2015 as a project at the Allen Institute for AI, a non-profit research institute founded in 2014. It is a semantic search engine based on an information retrieval algorithm powered by AI. It was created to highlight correlations between articles by studying their underlying semantics. It includes more than 200 million indexed academic papers and provides open-source APIs. It is possible to access the data in the dataset via queries that can be made in Python. The aim of Semantic Scholar is to build the Semantic Scholar Academic Graph (S2AG), which is a knowledge graph including information on papers, authors, and citations.

WGND 2.0 is an information resource that provides information about the gender probability of a person given her/his first name and country. It includes 3,922,294 names and 4,970,295 records. Each record is structured as a 4-tuple `(name, country_code, gender, gender_likelihood)`, including the first name, a country code (e.g., IT for Italy and FR for France), the gender (e.g, male, female, undetermined), and its likelihood. For each first name there can be one or more records. The limitation of WGND 2.0 is that although it is very extensive and complex, it does not include all names, in which case one must assign an undetermined value.

Gender Prediction Manager

The *Gender Prediction Manager* is in charge of four tasks. Initially, it coordinates interoperation activities with external modules, such as Semantic Scholar and WGND 2.0. Next, it focuses on extracting the first name of authors from their full name and the list of aliases. Subsequently, it assigns a probability to the simplified gender categorization adopted in this study (i.e., male, female, and undetermined). Finally, it determines the gender of an author based on the principle of fairness. Figure 2 shows the modules composing the *Gender Prediction Manager*: the *orchestration module*, the *discover first name module*, the *gender prediction module*, and the *module for fair gender prediction*.

The *Gender Prediction Manager* was developed using the Python programming language.

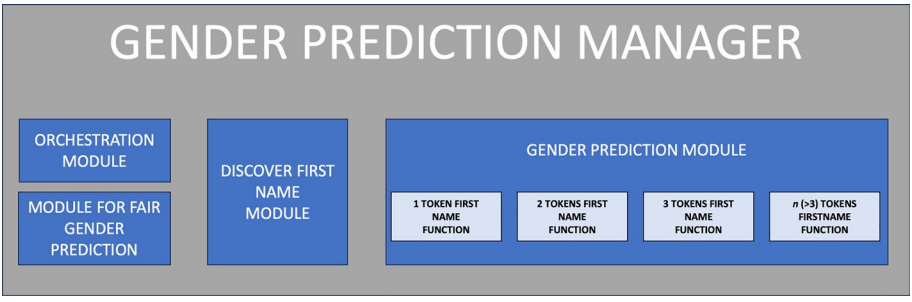

Fig. 2 Modules of the *Gender Prediction Manager*

Orchestration module

The *orchestration module* coordinates all the tasks performed by the software. It allows to carry out various tasks, such as access to WGND 2.0, communication with Semantic Scholar, and calls to the other mentioned modules. Algorithm 1 (see Appendix) outlines in pseudocode the sequence of steps aimed at retrieving authors' information from Semantic Scholar, beginning with the set of DOIs gathered from Scopus. The retrieved information encompasses the author's name, aliases, external IDs, and affiliations.

Discover first name module

The *discover first name module* aims to determine the primary name for gender prediction from the full name and the set of aliases. Initially, the module assumes that the last segment of the string represents the surname, while the remaining part serves as the potential first name candidate. Subsequently, the module calculates the number of tokens in the candidate and determines which token to exclude, such as dotted sections or any prefix of the surname and the subsequent portion.

Algorithm 2 (see Appendix) provides an explanation of how this functionality is achieved. For the sake of brevity, the algorithm omits the functions devoted to removing special characters from strings. Typically, the last token in the string, constituting the full name, is treated as the surname and excluded. Subsequently, the remaining string undergoes tokenization and is placed into a vector for analysis. If the vector contains only one element, and if that element consists of more than one character, it is regarded as the first name. For vectors with multiple elements, the following logic is applied to determine the first name. If all the elements in the vector are found together in WGND 2.0 (the name database used as reference), the entire vector is considered the first name. This applies even if the individual elements appear separately in WGND 2.0. Conversely, only the tokens present in WGND 2.0 are considered as the first name if they are not jointly present.

Gender prediction module

To categorize an author as female, male, or undetermined, we calculate the gender probability by utilizing WGND 2.0. The *gender prediction module* cross-references the previously determined first name to identify countries where the name is present. Next, it conducts calculations to estimate a new gender probability based on the gender probabilities within each country and, optionally, its population. We opted not to consider affiliation information due to research mobility. Researchers often relocate throughout their careers, and their native country seldom aligns with the country of their affiliated universities (Deville et al., 2014). The module is capable of handling various scenarios. When the first name consists of a single token, the algorithm aggregates gender probabilities across the countries where the name is present. It then calculates the average probability for each gender (i.e., male, female, undetermined) in one of two ways:

- "Countries" mode: the algorithm averages gender probabilities by simply counting the number of countries where the name appears, giving each country equal weight, regardless of population size.
- "Population" mode: the algorithm weighs each country's gender probability according to its population size, giving more influence to countries with larger populations.

If the first name contains two or three tokens, the module first attempts to find an exact match in the WGND 2.0 database. If a match is found, the gender probabilities are estimated as in the single-token case. If no exact match is found, the module examines each token of the name separately and aggregates the gender probabilities across countries where each token is present, again using either the "countries" mode or "population" mode to determine the final probabilities.

Finally, it computes the composite probability of each token by utilizing probability trees depicted in Fig. 3, representing all possible scenarios.

We denote the tokens comprising a first name as n-tuples, where each token is labeled using a combination of two components: a letter representing the gender prediction for that token (F for female, M for male, and U for undetermined), and a number indicating the token's position in the sequence (e.g., 1, 2, 3). Figure 3 illustrates the combinatorial structure for names composed of two (left side) and three (right side) tokens. For two-token names, this results in 9 possible combinations (3 gender classes × 3 positions). For three-token names, the total rises to 27 combinations (3^3). Rather than list all combinations explicitly, we represent them in a probability tree format to show how each token contributes to the overall gender prediction through its position and associated probability.

We calculate gender probabilities for first names consisting of two or three tokens using a straightforward combination of token-level predictions. These algorithms compute a composite probability, treating the genders of the individual tokens as independent events. For example, let's consider the probability when the first name is "Paul Maria". The probability of the gender being male is determined by multiplying the probability associated with the name "Paul" being male by the probability associated with "Maria" also being male. Algorithm 3 (see Appendix) presents pseudocode illustrating how to compute

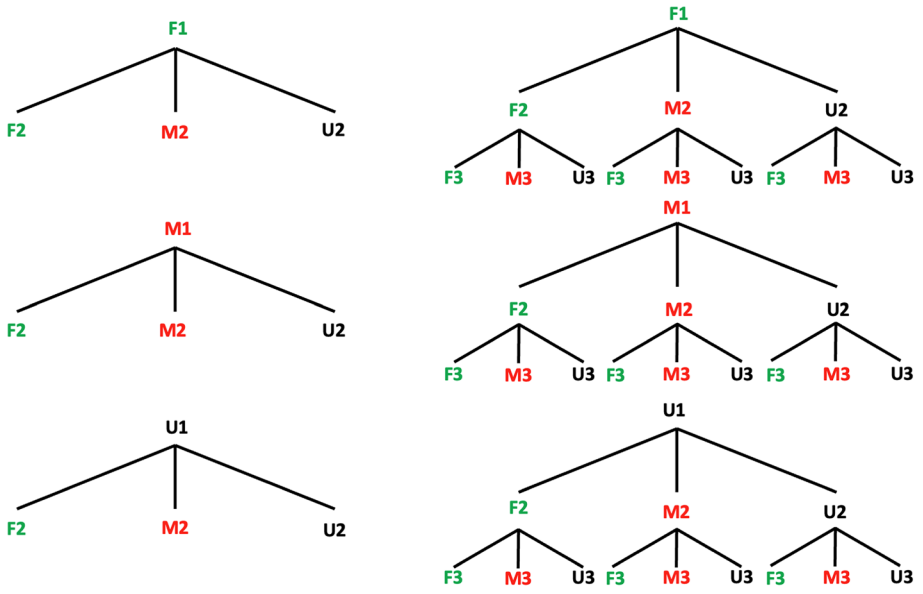

Fig. 3 Probability trees in case name of two tokens (on the left) and three tokens (on the right) first name. Each token is labelled as a combination of two parts: a character representing the gender (F for female, M for male, and U for undetermined) and a number (from 1 to 3) indicating the token's position in the sequence

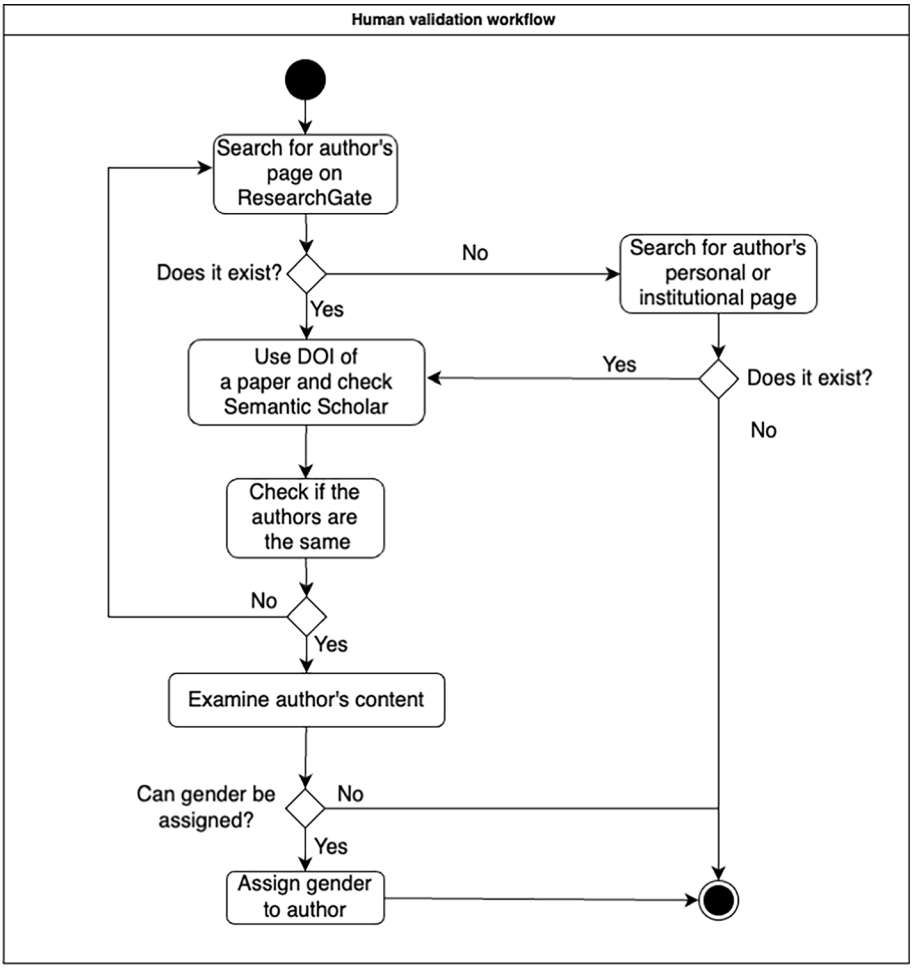

Fig. 4 Workflow for human validation of author's gender

compound probability for a first name composed of three tokens by weighting the countries' contribution by their population. The former assumption is just an Ansatz, since we know that in several languages the apposition of Maria after a masculine name is very common and it does not affect the gender, however that rule is not universal. Future versions will account for those features, more in deep. For brevity, the algorithm based on the number of countries is omitted. This variant can be regarded as a modification of algorithm 3, replacing the weighting factor with the likelihood that an author originates from a country, assumed to be equal for all countries. Here, the i represents the simplified gender classification adopted in this paper while j stands for one of the tokens.

Presently, the *Gender Prediction Manager* searches for exact matches in the Harvard dictionary for first names with more than three tokens, treating them as if they were first names consisting of only one token.

Module for fair gender prediction

The *module for fair gender prediction* makes the final decision on the gender to assign to an author. This is accomplished by using a threshold value, determined through an innovative semi-automated heuristic technique that we devised to ensure the fairness of name-based gender prediction. Accordingly, to assign a gender, the probability for that gender should exceed the threshold value.

Initially, we manually assign gender to a set of authors using the predefined workflow depicted in Fig. 4. Subsequently, we calculate precision ($\mathcal{P}$), recall ($\mathcal{R}$), and the F1-score to gauge the *Gender Prediction Manager*'s ability to recognize male and female authors. Precision represents the fraction of the relevant occurrences among all retrieved occurrences. Recall denotes the fraction of the relevant occurrences. F1-score represents the harmonic average of precision and recall, and provides a balanced representation of precision and recall in one metric. The formula for computing precision for each gender is $\mathcal{P}_i = \frac{TP_i}{TP_i+FP_i}$, recall for each gender is $\mathcal{R}_i = \frac{TP_i}{TP_i+FN_i}$, and F1-score is $\text{F1-score}_i = \frac{2 \cdot \mathcal{P}_i \cdot \mathcal{R}_i}{\mathcal{P}_i+\mathcal{R}_i}$, where TP_i, FP_i, and FN_i denote the number of true positives, false positives, and false negatives, respectively, for gender class i. Next, we analyze the ratio between the F1-score for females and males ($\text{F1-score ratio} = \frac{\text{F1-score}_F}{\text{F1-score}_M}$) across an increasing number of validated authors, plotting the F1-score ratio against a threshold value ranging from 0.5 to 1. The fairest solution would be to select the lower threshold value where the ratio is close to 1. Indeed, lower threshold values correspond to a higher number of authors with recognized gender but unequal results for males and females. To strike a balance between achieving a sufficiently high number of gender determinations and ensuring fairness across gender predictions, we adopt an approach that selects the threshold value which maximizes the harmonic mean of the F1-score ratio and accuracy $\mathcal{A}$ (i.e., the number of correctly predicted genders out of the total number of gender detection attempts). While this metric reflects our goal of jointly optimizing performance and fairness, we acknowledge that other multi-objective formulations could also be considered.

To determine the appropriate number of authors for computing performance metrics, we recommend calculating the ratio between the F1-scores for females and males across varying numbers of validated authors and graphing this ratio to identify the point at which it stabilizes. This point indicates the minimum number of validated authors required.

The human validation workflow, mentioned earlier, operates as follows (see Fig. 4). Initially, the author's page is searched for on ResearchGate, a social networking site for scientists. If the page is found, the DOI of one of the candidate author's papers is utilized as input for a query in Semantic Scholar. If the candidate and the author being validated share the same Semantic Scholar ID, we then examine author's content for keywords, such as pronouns (e.g., he, she), to help infer the gender. If found, we assign the gender. If the author's page on ResearchGate is not found, we then search for the author's personal or institutional page and repeat the process.

To avoid potential confusion regarding the use of the term "undetermined" in both algorithmic and human annotation contexts, we clarify our operational definition as follows: an author is labeled as "undetermined" either when the software cannot assign a gender due to name ambiguity or lack of data, or when human annotators cannot confidently infer gender due to insufficient publicly available information (such as the absence of a personal webpage or ResearchGate profile). While these two scenarios arise from distinct causes, we intentionally use the same term in both cases to ensure consistency between automated

Table 1 Gender rates by threshold values in case of gender prediction module operating in the population (pop.) and in the countries modes (cou.) for the energy transition dataset

Threshold	Females (pop.)	Males (pop.)	Undet. (pop.)	Females (cou.)	Males (cou.)	Undet. (cou.)
0.50	22.3%	45.5%	32.2%	25.6%	52.2%	22.2%
0.55	21.4%	43.7%	34.9%	24.9%	51.7%	23.4%
0.60	20.4%	40.7%	38.9%	24.5%	51.1%	24.4%
0.65	19.2%	38.1%	42.7%	23.9%	50.2%	25.9%
0.70	18.9%	37.5%	43.6%	23.4%	48.8%	27.8%
0.75	18.6%	35.8%	45.6%	22.3%	48.0%	29.7%
0.80	18.0%	35.2%	46.8%	21.7%	46.9%	31.4%
0.85	17.4%	34.3%	48.3%	21.1%	45.6%	33.3%
0.90	17.1%	33.9%	49.0%	20.5%	43.9%	35.6%
0.95	16.4%	33.2%	50.4%	19.5%	40.9%	39.6%

outputs and human-annotated ground truth. This approach facilitates direct and transparent comparison throughout our evaluation.

Experimentation

Datasets

The experimentation involves two distinct datasets. The first dataset consists of papers related to energy transition, while the second pertains to papers on critical infrastructures.

First, we retrieved relevant papers pertaining to energy transition from Scopus. For this purpose, the query included the terms “energy transition"OR “energy transformation". Initially, the query yielded 17,591 papers, which were subsequently narrowed down to 10,130 after filtering them using the Energy Systems Ontology (ESO) (De Nicola et al., 2024). These papers were written by 27,363 authors.

Next, we collected 380 papers published from 2006 to 2022 published in the proceedings of the International Conference on Critical Information Infrastructures Security (CRITIS), a leading event for practitioners and researchers in the field of critical infrastructure. In total, 929 authors were identified.

Research objectives

The research objectives of the experimentation were to predict the gender of the authors based on their names, to evaluate the quality of the *Gender Prediction Manager*, and to compare the quality of the results in terms of accuracy and gender fairness with other tools, namely, Gender API, ChatGPT, and Namsor.

Results

Therefore, as outlined in the earlier subsection titled *Module for Fair Gender Prediction*, we calculated precision, recall, and the F1-score to gauge the software’s effectiveness in predicting gender based on the first name. We evaluated two distinct software

Table 2 Gender rates by threshold values in case of gender prediction module operating in the population (pop.) and in the countries modes (cou.) for the critical infrastructures dataset

Threshold	Females (pop.)	Males (pop.)	Undet. (pop.)	Females (cou.)	Males (cou.)	Undet. (cou.)
0.50	16.15%	68.46%	15.39%	17.22%	74.6%	8.18%
0.55	15.39%	66.42%	18.19%	16.58%	74.49%	8.93%
0.60	14.64%	61.25%	24.11%	16.47%	74.38%	9.15%
0.65	14.21%	58.56%	27.23%	16.25%	73.30%	10.44%
0.70	13.35%	57.48%	29.17%	16.15%	72.87%	10.98%
0.75	13.13%	56.94%	29.92%	15.50%	72.01%	12.49%
0.80	12.59%	55.11%	32.29%	15.39%	70.94%	13.67%
0.85	12.27%	54.36%	33.37%	15.29%	69.86%	14.85%
0.90	12.06%	54.14%	33.8%	14.75%	68.25%	17.01%
0.95	11.52%	53.61%	34.88%	13.56%	65.88%	20.56%

Table 3 Number and gender of authors in the validated datasets

Dataset	Females	Males	Undet.	% Females	% Males	% Undet.
Energy transition	260	575	165	26.00%	57.50%	16.50%
Critical infrastructures	153	768	8	16.47%	82.67%	0.86%

configurations: one where the gender likelihood is computed solely based on the number of countries where a name is present (i.e., countries mode), and the other where the contributions of countries are weighted by their respective populations (i.e., population mode).

The number of predicted authors varies depending on the chosen probability threshold. Table 1 illustrates the percentage of detected males, females, and undetermined authors for different threshold values for the energy transition dataset. These percentages are shown when the prediction module operates in two modes: population (left side of the table) and countries (right side of the table). Similarly, Table 2 presents the percentage breakdown by threshold values for the authors in the validated critical infrastructures dataset. We observe that the number of undetermined authors increases as the threshold is augmented.

To verify the software's performance, a group of three individuals manually validated 1,000 randomly selected authors from the energy transition dataset and all the 929 from the critical infrastructures dataset. In both cases, they used the validation workflow (refer to Fig. 4) detailed in the subsection titled *Module for Fair Gender Prediction*. The 1000 energy transition validated authors comprise 260 females, 575 males, and 165 of undetermined gender, while the 929 critical infrastructures validated ones include 153 females, 768 males, and 8 of undetermined gender (see Table 3). To ensure transparency and reproducibility, all data used in this study, including the per-author prediction results from human validation and all automated tools are available at Guariglia et al. (2025).

To determine the optimal threshold probability, we computed accuracy of the *Gender Prediction Manager* and precision, recall, and the F1-score for both males and females among the 1000 validated authors (refer to the subsection titled *Module for Fair Gender Prediction*), while considering increasingly larger groups, starting from groups of 100, 200, and extending up to 1000. Figure 5 illustrates the case of the 1000 authors of the energy transition dataset when the gender prediction module operates in the countries mode while Fig. 6 in the population mode. For the sake of conciseness, we omit the graphs

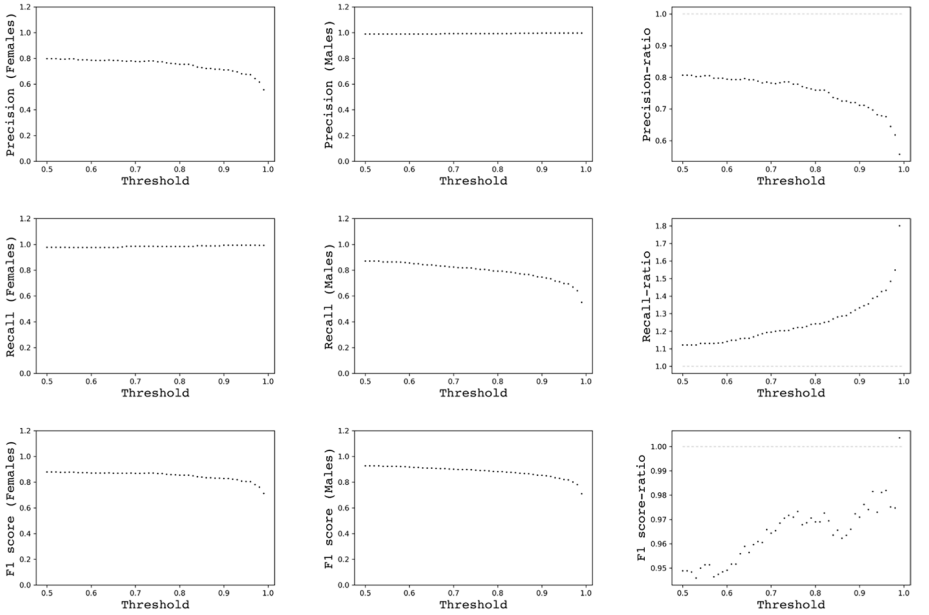

Fig. 5 Precision, recall, F1-score in case of 1000 validated authors and gender prediction module operating in the countries mode. The graphs are based on the energy transition dataset

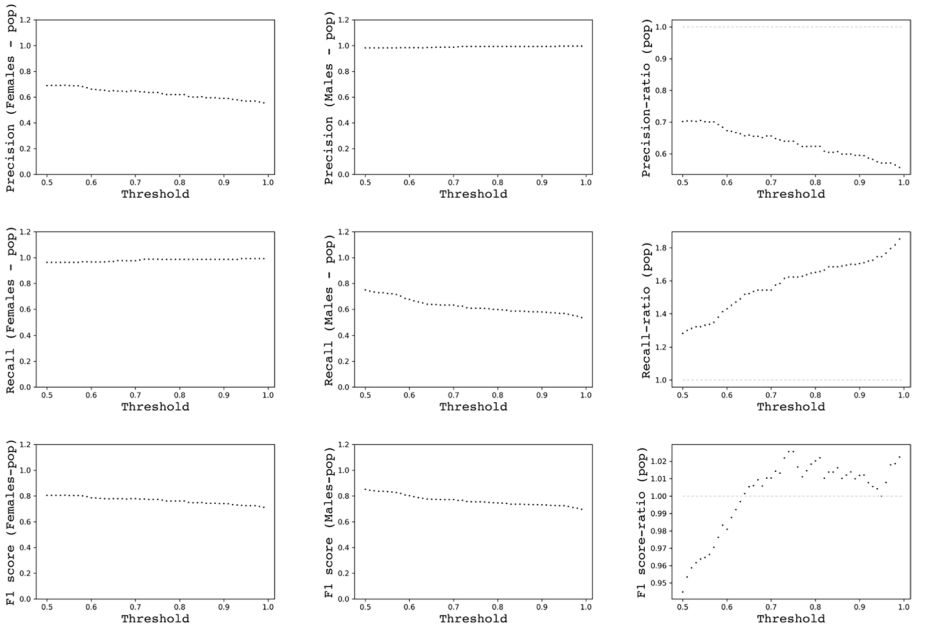

Fig. 6 Precision, recall, and F1-score in case of 1000 validated authors and gender prediction module operating in the population mode. The graphs are based on the energy transition dataset

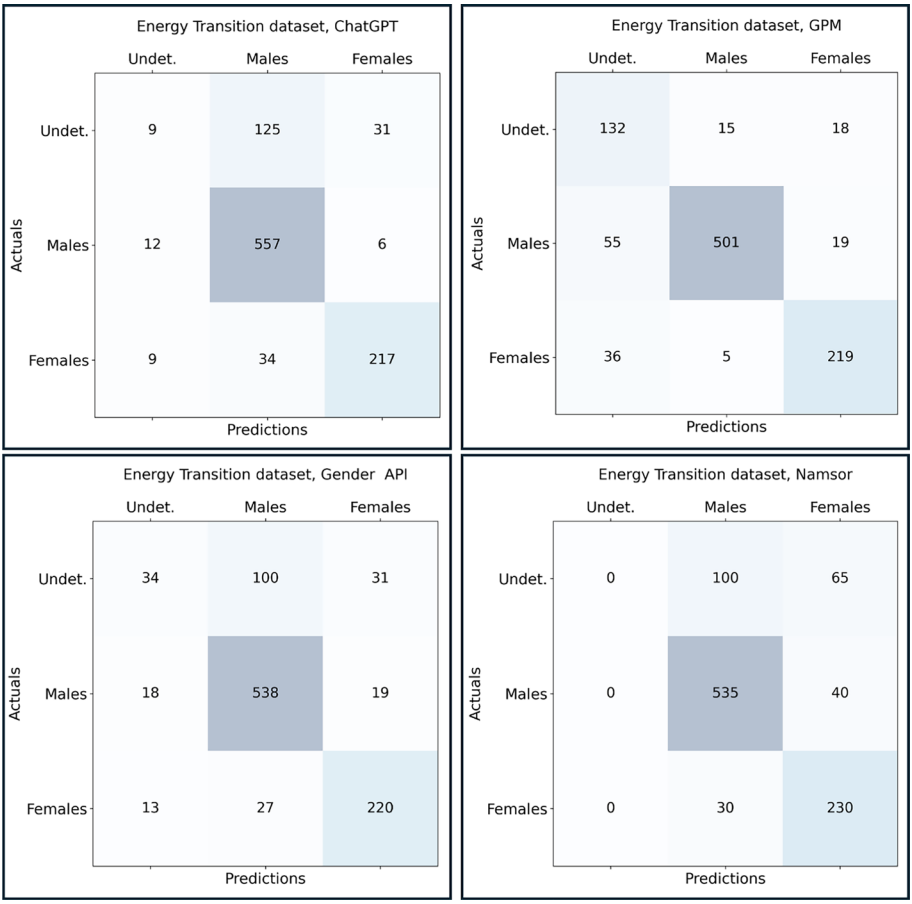

Fig. 7 Confusion matrices indicating the number of names correctly predicted, as well as the gender prediction errors made by ChatGPT, the *Gender Prediction Manager* (GPM), Gender API, and Namsor. Matrices are based on the energy transition dataset

based on the critical infrastructures dataset. As mentioned, the optimal threshold value is the one that maximizes the harmonic average of the F1-score ratio and accuracy $\mathcal{A}$. Table 4 reports the achieved results for the configurations of the *Gender Prediction Manager* for the energy transition dataset that produce the highest harmonic average of accuracy and F1-score ratio and the one that maximizes fairness. The optimal threshold probability value for the country mode is 0.51, while for the population mode is 0.50. It is worthy to mention that, even if selecting the threshold values that lead to F1-score ratio close to 1 would improve fairness of name-based gender prediction, maximizing the harmonic average of $\mathcal{A}$ and F1-score ratio is a good trade off between fairness and performance. Similarly, Table 5 reports the achieved results for the configurations of the *Gender Prediction Manager* that produce the highest harmonic average of $\mathcal{A}$ and F1-score ratio for the critical infrastructures dataset and the one that maximizes fairness. In such case, the optimal threshold probability value is 0.50 both for the country mode and for the populations mode.

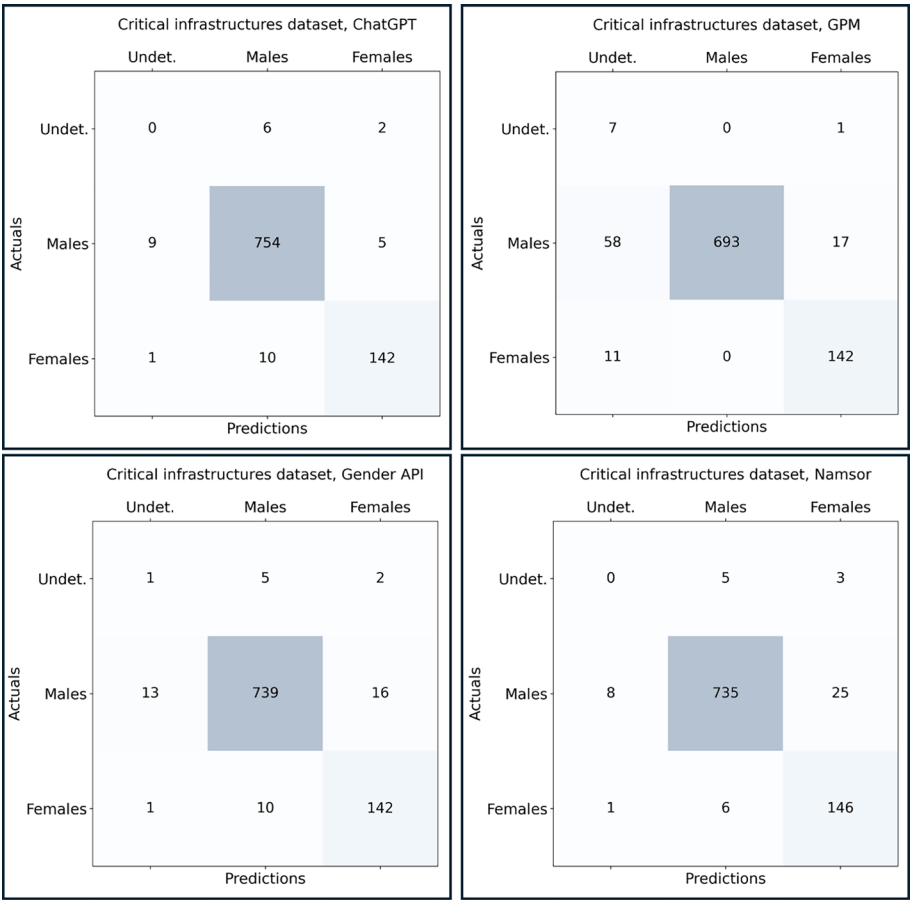

Fig. 8 Confusion matrices indicating the number of names correctly predicted, as well as the gender prediction errors made by ChatGPT, the *Gender Prediction Manager* (GPM), Gender API, and Namsor. Matrices are based on the critical infrastructures dataset

We also compared the performance of the *Gender Prediction Manager* with that of two of the most used commercial tools, i.e., *Gender API* and *Namsor*, and one of the most used large language models, namely ChatGPT[12]. To ensure a fair comparison, all gender prediction tools evaluated in this study received the same full names as input, as retrieved and expanded from Semantic Scholar. No tool was provided only with initials or partial names. This step was explicitly taken to avoid bias and ensure transparency and reproducibility of our results. Our study reveals that, for the energy transition case study (see Table 4), the *Gender Prediction Manager* operating in the countries mode outperforms all the other tools both for accuracy (0.85) and F1-score ratio (0.95). Conversely, the optimal solution considering only fairness for the energy transition dataset is the *Gender Prediction Manager* operating in the population mode with threshold 0.95 since the F1-score ratio is 0.98.

[12] ChatGPT website: https://chatgpt.com

Table 4 Comparison of performance metrics of the *Gender Prediction Manager* operating in the countries (cou.) and in the population (pop.) modes with Gender API, Namsor, and ChatGPT

Software	Mode	Threshold	$\mathcal{A}$	$F1_F$	$F1_M$	F1ratio	h_ave
Gender Prediction Manager	cou.	0.51	**0.85**	0.88	0.93	**0.95**	0.90
Gender Prediction Manager	pop.	0.50	0.76	0.81	0.85	0.94	0.84
Gender Prediction Manager	pop.	0.95	0.73	0.81	0.82	0.98	0.84
Gender API	-	-	0.79	0.88	0.94	0.93	0.85
Namsor	-	-	0.77	0.87	0.94	0.92	0.84
ChatGPT	-	-	0.78	0.90	0.96	0.94	0.85

The best values for accuracy (A) and the F1-score ratio are shown in bold

$\mathcal{A}$, $F1_F$, $F1_M$, F1-scoreratio, and h_ave indicate, respectively, accuracy, F1-scorefor females, F1-scorefor males, F1-scoreratio, and the harmonic average between $\mathcal{A}$ and F1-scoreratio. Results are based on the energy transition dataset

However, in this case, the accuracy is very low (0.73). Similarly, for the critical infrastructures dataset (see Table 5), the optimal solution for fairness is again the *Gender Prediction Manager* operating in the population mode with threshold 0.99 since the F1-score ratio is 1.00, while ChatGPT outperforms all the other tools for accuracy.

It is important to note that while our tool achieves competitive performance and, in some cases, enhanced fairness compared to commercial alternatives, it does not consistently outperform all baselines across all datasets. In particular, in the CRITIS dataset, commercial tools such as ChatGPT, Gender API, and NamSor demonstrated higher accuracy.

Figures 7 and 8 display the confusion matrices for the energy transition and critical infrastructures datasets, respectively. These matrices indicate the number of names correctly predicted, as well as the gender prediction errors made by each tool. Additionally,

Table 5 Comparison of performance metrics of the *Gender Prediction Manager* operating in the countries (cou.) and in the population (pop.) modes with Gender API, Namsor, and ChatGPT

Software	Mode	Threshold	$\mathcal{A}$	$F1_F$	$F1_M$	F1ratio	h_ave
Gender Prediction Manager	cou.	0.50	0.91	0.91	0.95	0.96	0.93
Gender Prediction Manager	pop.	0.50	0.81	0.79	0.90	0.89	0.85
Gender Prediction Manager	pop.	0.99	0.62	0.78	0.79	**1.00**	0.77
Gender API	-	-	0.95	0.91	0.95	0.95	0.95
Namsor	-	-	0.95	0.90	0.97	0.92	0.93
ChatGPT	-	-	**0.96**	0.95	0.99	0.96	0.96

The best values for accuracy (A) and the F1-score ratio are shown in bold

$\mathcal{A}$, $F1_F$, $F1_M$, F1ratio, and h_ave indicate, respectively, accuracy, F1-score for females, F1-score for males, F1-score ratio, and the harmonic average between $\mathcal{A}$ and F1-score ratio. Results are based on the critical infrastructures dataset

Table 6 Comparison of average accuracy ($\overline{\mathcal{A}}$) and F1 ratio (average) values of the *Gender Prediction Manager* operating in the countries (cou.) with the corresponding ones of Gender API, Namsor, and ChatGPT

Software	Mode	$\overline{\mathcal{A}}$	F1 ratio (average)
Gender Prediction Manager	cou.	**0.8792**	**0.9542**
Gender API	-	0.8707	0.9400
Namsor	-	0.8567	0.9221
ChatGPT	-	0.8737	0.9499

The best values are shown in bold

the matrices indicate that ChatGPT, Gender API, and Namsor tend to assign a gender even when it cannot be determined through human validation. This tendency may further reduce the overall quality of the results.

Table 6 compares the average results achieved by each tool across the two datasets, showing that, on average, the *Gender Prediction Manager* outperforms the others in both accuracy and F1-score ratio. While this table includes average values for ease of comparison, we emphasize that these are calculated over only two datasets, each with distinct distributions. Therefore, the main interpretation should be based on the detailed results provided for each dataset separately.

Statistical significance

We examined whether the differences in accuracy and F1-score ratio between the *Gender Prediction Manager*, in its optimal configuration maximizing the harmonic average between F1-score ratio and accuracy, and the other tools are statistically significant.

Regarding accuracy, we tested the following hypotheses for the *Gender Prediction Manager* (in its optimal configurations) in comparison with ChatGPT, Namsor, and Gender API. The null hypothesis (H_0) states that there is no significant difference in accuracy between the two software tools, meaning both have the same accuracy. In contrast, the alternative hypothesis (H_1) proposes that there is a significant difference in accuracy between the two tools. Since accuracy concerns the correct or incorrect classification of each software for each individual, we built a contingency table to compare the correct and incorrect classifications between the two software. The table includes the correct and incorrect classifications of software 1 and software 2. Since we are dealing with categorical data (correct/incorrect classifications), we used the McNemar test (McNemar, 1947) to compare the accuracies of the two tools. This test is designed to check whether there are significant differences in the performance of two classifiers in an experiment with paired samples (as in this case, where the same set of people is classified by two softwares). First,

Table 7 The χ^2 statistic and p-value obtained in the assessment of the statistical significance of accuracy

Dataset	Sw_1	Sw_2	χ^2	p-value
Energy transition	GPM	Gender API	6.25	0.01241
Energy transition	GPM	ChatGPT	9.04	0.00265
Energy transition	GPM	Namsor	17.64	0.00003
Critical infrastructures	GPM	Gender API	18.68	0.00002
Critical infrastructures	GPM	ChatGPT	28.31	0
Critical infrastructures	GPM	Namsor	17.98	0.00002

we computed $\chi^2 = \frac{(|b-c|-1)^2}{b+c}$, where b is the number of cases correctly classified by software 1 but not by software 2; c is the number of cases correctly classified by software 2 but not by software 1; -1 is a correction for continuity, used to avoid overestimation in small samples. Once we obtained the χ^2, we compared it with a critical value of the χ^2 distribution with 1 degree of freedom to determine the p-value. If the p-value were less than 0.05, we concluded that the difference in accuracy is statistically significant. Table 7 presents the results obtained, which indicate that the null hypothesis should be rejected in favor of the alternative hypothesis, i.e., there is a significant difference in accuracy between the *Gender Prediction Manager* and all the other tools.

Regarding the F1-score ratio, we tested the following hypotheses for the *Gender Prediction Manager* (in its optimal configurations) compared to ChatGPT, Namsor, and Gender API. The null hypothesis (H_0) asserts that there is no significant difference in the F1-score ratio between the two software tools, meaning they have the same F1-score ratio. In contrast, the alternative hypothesis (H_1) suggests that there is a significant difference in the F1-score ratio between the two tools.

To test if the differences in F1-score ratios are statistically significant, we employed bootstrapping. Specifically, we generated 10,000 bootstrap samples by sampling with replacement from the original datasets. For each bootstrap sample, we calculated the F1-scores for females and males for both software tools and computed the ratio for each software. We then obtained two bootstrap distributions of the ratios for the two tools. Next, we evaluated the difference between these distributions within the bootstrap iterations to assess whether the observed difference in the original sample was significant. Finally, we calculated the confidence intervals for the difference between the ratios.

Regarding multiple-comparisons correction, we conducted a limited number of planned, hypothesis-driven pairwise comparisons between our method and three baseline tools, across two datasets. Following common practice for such predefined comparisons, we did not apply a Bonferroni (Bonferroni, 1936) or other correction. However, we acknowledge that correction methods (e.g., Bonferroni) may be warranted in broader or exploratory studies and have made this decision explicit in the manuscript.

We found that the difference in F1-score ratios between the *Gender Prediction Manager* and Gender API was not statistically significant at the 95% confidence level for either the energy transition or the critical infrastructures dataset. Similarly, the difference in F1-score ratios between the *Gender Prediction Manager* and ChatGPT was not statistically significant at the 95% confidence level for either dataset. Lastly, the difference in F1-score ratios between the *Gender Prediction Manager* and Namsor was statistically significant at the 95% confidence level for the energy transition dataset, but not statistically significant at the 95% level for the critical infrastructures dataset.

In addition to statistical significance, we report effect sizes, i.e., Cohen's *h* (Cohen, 1988), for differences in accuracy and $\log_2$-transformed effect sizes (Julious, 2004) for fairness ratio comparisons. For the Energy Transition dataset, effect sizes range from 0.157 to 0.205, indicating small but consistent improvements of our system over baseline tools; for the CRITIS dataset, Cohen's *h* values range from –0.158 to –0.207, indicating small differences in favor of the comparison tools. These findings indicate that although some differences are statistically significant, their practical impact is modest.

To further assess practical differences in gender fairness, we computed $\log_2$-transformed effect sizes based on the F1-score ratio for female versus male classification ($F1_F/F1_M$). For the Energy Transition dataset, these differences between our *Gender Prediction Manager* and the comparison tools are 0.015 (vs. ChatGPT), 0.031 (vs. Gender API), and 0.046 (vs.

Table 8 Accuracy of gender prediction tools for the Energy Transition dataset

Tool	Accuracy (case 1)	Accuracy (case 2)
Gender Prediction Manager	0.87	0.97
ChatGPT	0.93	0.96
Gender API	0.91	0.96
NamSor	0.91	0.94

Case 1: accuracy only for validated authors who are not undetermined. Case 2: accuracy calculated only for those validated authors who are assigned a determinate gender (not"undetermined") by both human validation and by all tools

Table 9 Accuracy of gender prediction tools for the CRITIS dataset

Tool	Accuracy (case 1)	Accuracy (case 2)
Gender Prediction Manager	0.91	0.98
ChatGPT	0.97	0.99
Gender API	0.96	0.98
NamSor	0.96	0.98

See caption of Table 8 for explanation of case 1 and case 2

NamSor), suggesting slightly more balanced gender performance by our tool. The CRITIS dataset shows $\log_2$-transformed effect sizes of 0.015 (vs. Gender API), 0.0 (vs. ChatGPT), and 0.061 (vs. NamSor).

Discussion

The research objectives of the presented experimentation were to demonstrate that the proposed software can predict the gender of authors based solely on their names only, to assess the performance of the *Gender Prediction Manager*, and to compare it to other tools used for name-based gender prediction. To this purpose, we used accuracy, F1-score ratio, to measure performance and fairness of the software in recognizing gender from names, and their harmonic average. Considering both the case studies, our study reveals that the *Gender Prediction Manager*, operating in the countries mode, on average, has the highest accuracy and F1-score ratio and achieves performance in the range of Gender API, NameSor, and ChatGPT. Despite that, the software operating in this mode does not represent the best solution for fairness. Indeed, the *Gender Prediction Manager* operating in the population mode is the fairest software. Unfortunately, the accuracy of the software degrades in the optimal fairness conditions (F1-score ratio close to 1). Therefore, a compromise is always required.

Several commercial tools, such as Gender API and NamSor, do provide probabilities and confidence scores that could, in principle, be thresholded to yield an “undetermined” label. However, our aim in this study was to evaluate these tools as they are typically used in practice, that is, according to their default configurations. This approach reflects

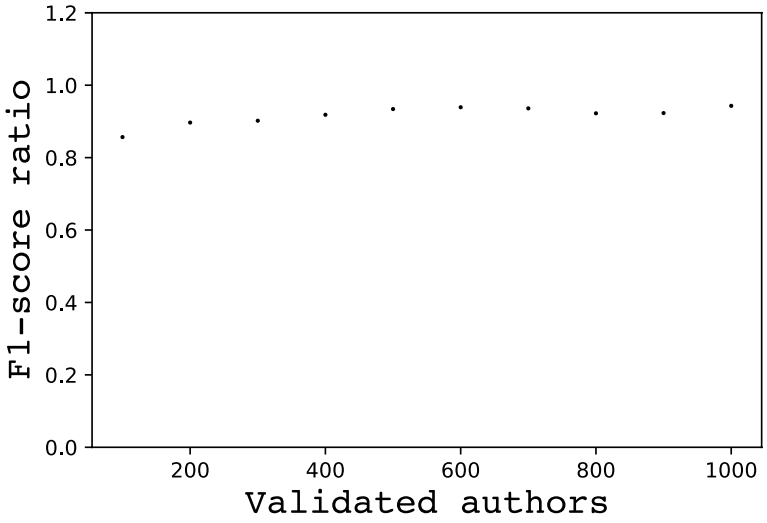

Fig. 9 Stability of F1-score ratio with respect to validated authors in the case of *Gender Prediction Manager* operating in the country mode

real-world scenarios, where most users rely on the standard, out-of-the-box behavior of such software. In contrast, our tool was designed with thresholding as a core feature, allowing users to explicitly manage trade-offs between accuracy and fairness, and to support the "undetermined"label as a first-class outcome. While we recognize that this introduces a certain asymmetry in the comparison, our intention was not to disadvantage the commercial tools. Rather, we aimed to highlight how the incorporation of fairness-aware thresholding mechanisms can influence performance and outcomes in practical applications. Nonetheless, we acknowledge that applying similar thresholding strategies to commercial tools could provide a more direct comparison in terms of methodological parity. Exploring the impact of post-processing thresholds across different tools, including commercial ones, is a compelling direction for future research.

Another important methodological consideration concerns how "undetermined" labels are treated in the evaluation process. As shown by our additional analyses, excluding cases labeled as "undetermined" in the ground truth, or restricting accuracy calculations to authors not labeled "undetermined" by any tool, results in substantially higher accuracy values across all methods, and in some cases changes the comparative ranking among tools. For instance, when undetermined cases are excluded, ChatGPT and commercial tools show higher apparent accuracy, while our tool's accuracy also increases substantially. These findings demonstrate that the evaluation protocol, specifically, the inclusion or exclusion of undetermined cases, can significantly influence reported outcomes and the conclusions drawn about relative tool performance. We emphasize the need for transparent reporting of evaluation strategies, and include these sets of results (see Table 8 and 9) to provide a more comprehensive perspective on the trade-offs involved in gender prediction tasks.

Our use of an "undetermined" class for low-confidence cases provides greater transparency and allows users to control the trade-off between fairness and completeness of gender assignments, a feature not typically available in commercial tools.

There are some threats to validity that could cause a degraded quality of results. Bias occurs when datasets are not managed correctly or they are treated differently, leading to

distortions in the resulting analysis and conclusions. This can happen in several reasons: incorrect data collection, inaccurate pre-processing, non-optimal algorithms, and misinterpreting results.

An example of incorrect data collection is when the names to be analyzed are incorrect. To mitigate this problem, we started the analysis from data extracted from Scopus, which is a widely known and highly reliable source of bibliometric information. Another example concerns, the size of the validated sample of authors. In order to assess whether 1000 authors were sufficient for determining the threshold probability in the experiment, we examined the stability of the F1-score ratio. Figure 9 shows the case of the *Gender Prediction Manager* operating in the country mode. In such case, the curve of the F1-score ratio remains consistent as the number of validated authors increases. Therefore, we can infer that the set of 1000 authors is adequate, and there is no need for further manual validation.

Inaccurate pre-processing can also introduce bias, mainly due to missed data cleaning and transformation before analysis. In our experimentation, overall, the F1-score ratio is typically below 1. An exception occurs when the *Gender Prediction Manager* operates in the population mode above the optimal threshold (>0.65), where the accuracy of the software, however, degrades. Since this is also observed in the case of Gender API, NameSor, and ChatGPT which likely employ different algorithms, the bias is likely inherent in the source data where the associations between names and genders are derived. This implies that a fairness adjustment should be applied using preprocessing methods (Buyl & de Bie, 2024).

Non-optimal algorithms are another threat for the success of our experimentation. Certain analysis techniques might be inherently prone to bias. For instance, the achieved results suggest that, although considering population might appear to be a more accurate model, the number of inhabitants of a country does not necessarily increase the likelihood that an author comes from that country. There are likely other factors, such as the country's income level, that should be taken into account. For this reason, we opted for two different configurations of the *Gender Prediction Manager* to optimize either the trade-off between accuracy and F1-score ratio (and, hence, fairness) of the software or only its fairness.

Finally, the last concern involves the potential misinterpretation of the results, which could lead to drawing incorrect conclusions. This could be due to lack of ground truth. For instance, if the precision in determining the female gender is very low, without ground truth, we might draw incorrect conclusions about the gender gap. To mitigate this issue, we engaged three individuals to manually determine the gender of authors. Then, to further address this, we considered several quality measures beyond accuracy, including precision, recall, and the F1-score, as well as their ratios, measured for males and females. Thus, we were able to evaluate both the software's ability to assign a gender to an author and its fairness in operation. While this manual validation step ensured reliability, it is time-intensive and limits scalability. Future work may explore automated methods, such as using NLP to extract gender cues from academic bios or web content, to reduce the manual burden.

We also recognize that while Semantic Scholar offers open API access and provides expanded name aliases, features that enhance prediction accuracy, its proprietary methods for compiling and disambiguating author data introduce a degree of opacity. This is an important trade-off: although our current implementation relies on Semantic Scholar for its accessibility and breadth, our pipeline is explicitly modular. Other metadata sources (e.g., Scopus, or institutional records) can be used as needed, without impacting the overall workflow or architecture. This flexibility allows researchers to tailor the pipeline to their transparency or data provenance requirements.

A practical limitation of our threshold-based procedure is its inherent dataset-specificity. The optimal threshold for fair gender assignment may vary across datasets due to differences in gender distributions and cultural naming conventions. While in our experiments the thresholds for the two case studies were similar, we emphasize that, in general, a new calibration process is required for each dataset to ensure fairness and optimal performance. This necessity prevents the tool from being fully plug-and-play, as it introduces an extra setup step before deployment. Although this calibration enables more precise and fair results, it does so at the cost of increased effort for users. Looking ahead, future work could explore methods to automate or simplify the threshold selection process, thereby improving the usability and accessibility of fairness-aware gender prediction in diverse research contexts.

An additional promising direction for future work involves the use of separate threshold values for male and female gender assignments. While our current approach employs a unified, fairness-aware threshold determined by the F1-score ratio, implementing gender-specific thresholds could offer finer control over performance and fairness. This could be especially valuable in datasets with asymmetric gender distributions or differing name ambiguities across genders. By optimizing thresholds for each gender independently, it may be possible to further reduce bias and improve the overall balance of gender prediction. We highlight this as a potential extension of our method and an area for future investigation.

Conclusion

Name-based gender prediction is crucial for studying gender diversity and, in particular, to determine whether a gender gap exists in a scientific community. While performing name-based gender prediction, the issue of fairness, i.e., equal capability for both genders, is central since erroneous assignment could introduce a bias that could eventually affect the quality of the gendered analysis. For the former reason, we have developed some software for name-based gender prediction and set up an innovative method capable of measuring the degree of fairness in its operation. The achieved results (upon a real test case) show that the performance of the software is satisfactory for both the dimensions we considered. Furthermore, the *Gender Prediction Manager* achieves performance in the range of Gender API, NameSor, and ChatGPT. In summary, while our tool does not universally outperform proprietary commercial tools, it provides a competitive, transparent, and fairness-aware approach to name-based gender prediction. The use of open data and tunable thresholds allows users to adapt the method to specific fairness and uncertainty goals, supporting more responsible and reproducible gender analyses in scientific studies.

It is important to note the differences between the two datasets: the energy transition dataset contains a higher percentage of female names and more Asian names, while the critical infrastructure dataset has a lower percentage of female names and more European names. These differences explain the lower accuracy observed with the energy transition dataset and highlight the contribution of the *Gender Prediction Manager* in addressing the challenges associated with predicting Asian names.

In principle, there are several directions to explore for further enhancement of software performances the software. These include utilizing additional datasets associating names with genders; and improving the tool's ability to estimate the probability of an author's origin country. All those possibilities represent ongoing works, however the elective solution would be to ask authors their gender and include the information on available datasets.

Appendix

Algorithm 1 Algorithm for collecting authors' information

```
doiSet = set()
paperDict = {}
authorsDict = {}
sch = SemanticScholar()
for all doi ∈ doiSet do
    paper = sch.get_paper(x)
    for all sch_author ∈ paper.authors do
        a_Id=y.authorId
        if a_Id ∉ authorsDict.keys() then
            a_Name=y.name
            a_Aliases=y.aliases
            a_ExtIds =y.externalIds
            a_Affiliations =y.affiliations
            author = Autore(a_Id, a_Name, a_Aliases, a_ExtIds,
a_Affiliations)
            authorsDict[a_Id] = author
        end if
    end for
end for
return authorsDict
```

Algorithm 2 Algorithm to choose which parts of name use for the prediction process

```
all_names_set = set()
long_name_vect = long_name.split()
returned_Name = null
if len(long_name_vect) > 1 then
    surname = long_name_vect[-1]
    firstname_vett=long_name_vect[:-1]
    token_1 = firstname_vett[0].lower()
    if len(firstname_vett) == 1 then
        if checkIfInitials(token_1) == False then
            returned_Name = token_1
        end if
    end if
    if (len(firstname_vett) > 1) then
        for i ∈ range(0, len(firstname_vett) do
            if token[i] ∈ all_names_set then
                if i == 0 then
                    returned_Name = token[1]
                else returned_Name = returned_Name + ‘‘ ’’ + token[i]
                end if
            end if
        end for
    end if
end if
return returned_Name
```

Algorithm 3 Algorithm for computing the gender probability (based on population) for first names with three tokens individually listed in the Harvard Dictionary

```
firstName = firstName_instance
token_vett = firstName.split()
i = [f, m, u]
prob_i = 0
probPop_i = 0
harvard.dict = {}
j = [1, 2, 3]
probPop_i,j = 0
for all (token ∈ token.vett) do
    inhabitants = 0
    countries_set = set()
    record_set_j = harvard.dict[token]
    for (loc_record ∈ record_set_j) do
        country = loc_record.countrycode
        pop = float(loc_record.population)
        prob = float(loc_record.prob)
        gender = loc_record.gender
        if country ∉ countries_set then
            countries_set.add(country)
            inhabitants = inhabitants + pop
        end if
    end for
    for all (loc_record ∈ record_set_j) do
        pop = float(loc_record.population)
        prob = float(loc_record.prob)
        gender = loc_record.gender
        if (gender == i) then
            probPop_i,j = probPop_i,j + prob·pop
        end if
    end for
    probPop_i,j = (1/inhabitants) · probPop_i,j
end for
probPop_f = probPop_f,1 · probPop_f,2 · probPop_f,3
probPop_m = probPop_m,1 · probPop_m,2 · probPop_m,3
probPop_u = 1 - probPop_f - probPop_m
return probPop_i,j
```

Acknowledgements We gratefully acknowledge the partial support of the gEneSys (Transforming Gendered Interrelations of Power and Inequalities in Transition Pathways to Sustainable Energy Systems) project, which has received funding from the European Union's Horizon Europe - Culture, creativity and inclusive society - under grant agreement no. 101094326.

Funding Open access funding provided by Ente per le Nuove Tecnologie, l'Energia e l'Ambiente within the CRUI-CARE Agreement.

Declarations

Conflict of interest All authors involved in this research have declared no Conflict of interest.

References

Abramo, G., Aksnes, D. W., & D'Angelo, C. A. (2021). Gender differences in research performance within and between countries: Italy vs Norway. *Journal of Informetrics, 15*(2), 101144. https://doi.org/10.1016/j.joi.2021.101144

Alford, R. D. (1987) Naming and identity: A cross cultural study of personal naming practices. Retrieved from, https://api.semanticscholar.org/CorpusID:141787656

Bèrubè, N., Ghiasi, G., Sainte-Marie, M., & Lariviere, V. (2020) Wiki-gendersort: Automatic gender detection using first names in wikipedia

Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del Regio Istituto Superiore di Scienze Economiche e Commericiali di Firenze, 8*, 3–62.

Bonham, K. S., & Stefan, M. I. (2017). Women are underrepresentad in computational biology: An analysis of the scholarly literature in biology, computer science and computational biology. *PLoS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1005134

Buyl, M., & Bie, T. (2024). Inherent limitations of AI fairness. *Communications of the ACM, 67*(2), 48–55. https://doi.org/10.1145/3624700

Choji, T., Moral-Munoz, J., & Cobo, M. (2024). Is the scientific impact of the LIS themes gender-biased? A bibliometric analysis of the evolution, scientific impact, and relative contribution by gender from 2007 to 2022. *Scientometrics*. https://doi.org/10.1007/s11192-024-05005-3

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Lawrence Erlbaum Associates.

De Nicola, A., & D'Agostino, G. (2021). Assessment of gender divide in scientific communities. *Scientometrics, 126*(5), 3807–3840. https://doi.org/10.1007/s11192-021-03885-3

De Nicola, A., Patriarca, T., Fresilli, B., Opromolla, A., Guariglia Migliore, M., Leonardi, N., D'Agostino, G., Cellini, M., Mirenda, C., Tagliacozzo, S., Pisacane, L., & Vassillo, C. (2024) D.1.2 - Report on gendered assessment of the energy systems knowledge community and EU policies for sustainable energy systems—Horizon Europe Project gEneSys—Transforming gendered interrelations of power and inequalities in transition pathways to sustainable energy systems, grant agreement no. 101094326. https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e509765b4f&appId=PPGMS

Deville, P., Wang, D., Sinatra, R., Song, C., Blondel, V. D., & Barabási, A.-L. (2014). Career on the move: Geography, stratification, and scientific impact. *Scientific Reports*. https://doi.org/10.1038/srep04770

Eagly, A., Nater, C., Miller, D., Kaufmann, M., Sczesny, S. (2019). Gender stereotypes have changed: A cross-temporal meta-analysis of us public opinion polls from 1946 to 2018. *American Psychologist*. https://doi.org/10.1037/amp0000494

Gautam, V., Subramonian, A., Lauscher, A., & Keyes, O. (2024). Stop! in the name of flaws: Disentangling personal names and sociodemographic attributes in NLP. In A. Faleńska, C. Basta, M. Costa-Jussà, S. Goldfarb-Tarrant, & D. Nozza (Eds.), *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)* (pp. 323–337). Association for Computational Linguistics.

Gomide, J., & F.D., & Kling H,. (2017). Name usage pattern in the synonym ambiguity problem in bibliographic data. *Scientometrics, 112*, 747.

Guariglia Migliore, M., D'Agostino, G., Patriarca, T., & De Nicola, A. (2025). Datasets for Fair Name-Based Gender Prediction in Scientific Communities. figshare. Dataset. https://doi.org/10.6084/m9.figshare.29909603.v1

Larivière, V., Ni, C., Gingras, Y., Cronin, B., & Sugimoto, C. R. (2013). Bibliometrics: Global gender disparities in science. *Nature, 504*(7479), 211–213.

Hough, C. (2016). *The Oxford Handbook of Names and Naming*. Oxford University Press.

Huang, J., Gates, A. J., Sinatra, R., & Barabási, A.-L. (2020). Historical comparison of gender inequality in scientific careers across countries and disciplines. *Proceedings of the National Academy of Sciences of the United States of America, 117*(9), 4609–4616. https://doi.org/10.1073/pnas.1914221117

Julious, S. A. (2004). *Sample Sizes for Clinical Trials*. Chapman & Hall/CRC.

Knowles, R., Carroll, J., & Dredze, M. (2016). Demographer: Extremely simple name demographics. In D. Bamman, A. S. Doğruöz, J. Eisenstein, D. Hovy, D. Jurgens, B. O'Connor, A. Oh, O. Tsur, & S. Volkova (Eds.), *Proceedings of the First Workshop on NLP and Computational Social Science* (pp. 108–113). Association for Computational Linguistics.

LGBTQIA Resource Center Lesbian, Gay, Bisexual, Transgender, Queer, Intersex, Asexual. Retrieved March 24, 2024, from https://lgbtqia.ucdavis.edu

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika, 12*(2), 153–157. https://doi.org/10.1007/bf02295996

Meganathan, R. (2009). The politics of naming: The search for linguistic and ethnic identity in Tamil Nadu. *Contributions to Indian Sociology, 43*(2), 317–324. https://doi.org/10.1177/006996670904300205

Payton, F. C., & Berki, E. (2019). Countering the negative image of women in computing. *Communications of the ACM, 62*(5), 56–63. https://doi.org/10.1145/3319422

Raffo, J. (2021). WGND 2.0. https://doi.org/10.7910/DVN/MSEGSJ

Sánchez-Jiménez, R., Guerrero-Castillo, P., Guerrero-Bote, V. P., Halevi, G., & De-Moya-Anegón, F. (2024). Analysis of the distribution of authorship by gender in scientific output: A global perspective. *Journal of Informetrics, 18*(3), 101556. https://doi.org/10.1016/j.joi.2024.101556

Santamaría, L., & Mihaljević, H. (2021). Comparison and benchmark of name-to-gender inference services. *PeerJ Computer Science*. https://doi.org/10.7717/peerj-cs.156

Sebo, P. (2021). Performance of gender detection tools: A comparative study of name-to-gender inference services. *Journal of the Medical Library Association, 109*, 414.

United Nations - Department of Economic and Social Affairs Sustainable Development. (2015). Transforming our world: The 2030 agenda for sustainable development. *Journal of Public Health, 37*, 13.

Van Buskirk, I., Clauset, A., & Larremore, D. B. (2023) An open-source cultural consensus approach to name-based gender classification. In: Proceedings of the Seventeenth International AAAI Conference on Web and Social Media (ICWSM2023)

Zhang, L., Sivertsen, G., Du, H., Huang, Y., & Glänzel, W. (2021). Gender differences in the aims and impacts of research. *Scientometrics, 126*(11), 8861–8886. https://doi.org/10.1007/s11192-021-04171-y

Authors and Affiliations

Maria Guariglia Migliore[1,2,3] · Gregorio D'Agostino[3] · Tatiana Patriarca[3] · Antonio De Nicola[3]

✉ Antonio De Nicola
antonio.denicola@enea.it

Maria Guariglia Migliore
maria.guarigliamigliore@uniroma1.it

Gregorio D'Agostino
gregorio.dagostino@enea.it

Tatiana Patriarca
tatiana.patriarca@enea.it

[1] Department of Engineering Science, Guglielmo Marconi University, Via Plinio 44, 00198 Rome, Italy

[2] Sapienza University of Rome, Department of Mechanical and Aerospace Engineering, Via Eudossiana, 18, 00184 Rome, Italy

[3] Italian National Agency for New Technologies, Energy and Sustainable Economic Development (ENEA), Casaccia Research Centre, Via Anguillarese 301, 00123 Rome, Italy