Detection and Analysis of Gender Bias in Media through NLP Techniques

Detección y análisis de sesgos de género en medios de comunicación a través de técnicas de PLN

M. Dolores Molina-González,[1] **Manuel Carlos Díaz-Galiano,**[1]
Isabel Cabrera-de-Castro,[1] **Sergio Maruenda-Bataller,**[2]
Maite Taboada,[3] **María Teresa Martín-Valdivia,**[1]
[1]CEATIC. Department of Computer Science, Universidad de Jaén, 23071, Jaén, Spain
[2] Departament de Filologia Anglesa i Alemanya, Facultat de Filologia,
Traducció i Comunicació, Universitat de València, Valencia, Spain
[3]Discourse Processing Lab, Department of Linguistics, Simon Fraser University,
Burnaby, Canada
[1] {mdmolina, mcdiaz, iccastro, maite}@ujaen.es
[2] {sergio.maruenda}@uv.es
[3] {mtaboada}@sfu.ca

Abstract: Gender inequality remains a pervasive global issue, with the struggle for women's rights gaining particular momentum in the latter half of the 20th century and into this century. The *Global Gender Gap Report* evaluates gender parity across several dimensions. Despite progress, no country has achieved full gender parity. However, Spain is actively pursuing equality in key sectors. Media representation is crucial for normalising female leadership and recognising women's contributions. Against this backdrop, this article presents a computational system to analyse gender representation in the Spanish press using Natural Language Processing (NLP) techniques, such as Named Entity Recognition (NER), quote extraction and coreference resolution. The system performs web scraping, identifies quotations, links them to speakers and predicts their gender to measure current representation. With a methodological focus on text pre-processing and quotation extraction, the findings highlight the usefulness of NLP for large-scale analysis, supporting efforts towards gender equality.
Keywords: Gender inequality, Name Entity Recognition, Media analysis, Coreference resolution

Resumen: La desigualdad de género persiste globalmente y aunque la lucha por los derechos de las mujeres ha ganado impulso, evidenciado por informes como *Global Gender Gap Report*, aún no se alcanza totalmente en ningún país. España busca activamente la igualdad en sectores clave. La representación mediática es crucial para normalizar el liderazgo femenino y reconocer sus contribuciones. En este contexto, este artículo presenta un sistema computacional para analizar la representación de género en la prensa española utilizando técnicas de Procesamiento del Lenguaje Natural (NLP), como el reconocimiento de entidades nombradas (NER), extracción de citas y resolución de correferencias. El sistema realiza web scraping, identifica citas, las vincula a oradores y predice su género para medir la representación actual. Con un enfoque metodológico en preprocesamiento de texto y extracción de citas, los hallazgos subrayan la utilidad del NLP para análisis masivos, apoyando los esfuerzos hacia la equidad de género.
Palabras clave: Desigualdad de género, reconocimiento de entidades nombradas, análisis de medios de comunicación, análisis de correferencia

1 Introduction

The lack of gender equality is one of the main forms of discrimination in the world. In this sense, the struggle for women's rights became one of the great movements of the 20th century and continues into this century (Gordon, 2002; Schuster, 2017). A recent iteration of the Global Gender Gap Report (GGGR)[1] evaluates the current state and progress towards gender equality across four key dimensions or subindexes (Economic Participation and Opportunity, Educational Attainment, Health and Survival, and Political Empowerment), showing that the overall gender gap score in 2025 for the 148 countries included in the latest edition stands at 68.8%, where 100% would represent full parity (Figure 1). No country has yet reached full gender parity.[2]

While the global picture of gender equality remains incomplete and the United Nations works towards gender equality across the world as part of the Sustainable Development Goals,[3] advanced societies are making demonstrable strides. According to the 2024 GGGR, Spain ranks 10th out of 146 countries, with a score of 0.797 (with 1 corresponding to full gender parity).

Nevertheless, the representation of women in most areas of global society is clearly lower than that of men, from political and executive leadership of a country[4] to professionals working in most areas of daily life through the scarcity of female leadership[5] or referents in science (Berenbaum, 2019).

Gender equality is one of the 17 Sustainable Development Goals set by the United Nations for the 2030 Agenda. While some noteworthy progress has been achieved, significant challenges persist. These challenges demand increased focus on increasing women's participation in public and professional spheres, alongside efforts to enhance their visibility across all domains.

Research suggests a persistent gender disparity within the media landscape. Women remain underrepresented in news writing, often failing to occupy roles as experts or protagonists within news narratives compared to their male counterparts.

To address this problem, the Global Media Monitoring Project (GMMP)[6] emerged as a significant international research initiative. The project stands as the most comprehensive study on gender representation in media on a global scale. Its core objective centers on advocating for a more balanced portrayal of women within media narratives. Every five years since 1995, the GMMP collects data on gender indicators in the news, such as the presence of women, gender bias and stereotypes. The last data collected through the GMMP for the year 2025 found that only 26% of those seen, read or heard in the media are women. In addition, the discourse advanced by the media tends to perpetuate gender stereotypes. In fact, it is also evident that only 9% of news stories feature stories that reflect gender inequality and even fewer, 4%, aim to challenge these stereotypes (Kassova, 2023; Kassova, 2020).

These numbers clearly stand to improve if we want to reach parity, as it is essential for women to have their own voice in the media and to be recognized as experts and references in all types of news.

Our project **GeMeCo**: *Visibilización de Brechas de **G**énero en **M**edios de **C**omunicación utilizando Inteligencia Artificial* (Visibility of Gender Gaps in Media using Artificial Intelligence) aims to make this problem visible as a first step towards achieving a more balanced representation. The main objective is to develop a dashboard capable of compiling news items published daily in various Spanish media outlets and then automatically identifying gender-related quotes and mentions, thereby creating an online repository of daily summaries that highlight indicators of gender gaps. However, several challenges hinder the automation of gender representation analysis in news media. These include the difficulty of reliably extracting both the direct and indirect quotations in Spanish, limitations in co-reference resolution tools for speaker attribution, and the complexity of gender prediction. Addressing these challenges is essential to de-

[1] https://www.weforum.org/publications/global-gender-gap-report-2025/

[2] https://www3.weforum.org/docs/WEF_GGGR_2025.pdf

[3] https://unstats.un.org/sdgs/report/2024/The-Sustainable-Development-Goals-Report-2024.pdf

[4] https://data.oecd.org/inequality/women-in-politics.htm

[5] https://www.pewsocialtrends.org/2018/09/20/women-and-leadership-2018/

[6] https://waccglobal.org/our-work/global-media-monitoring-project-gmmp/

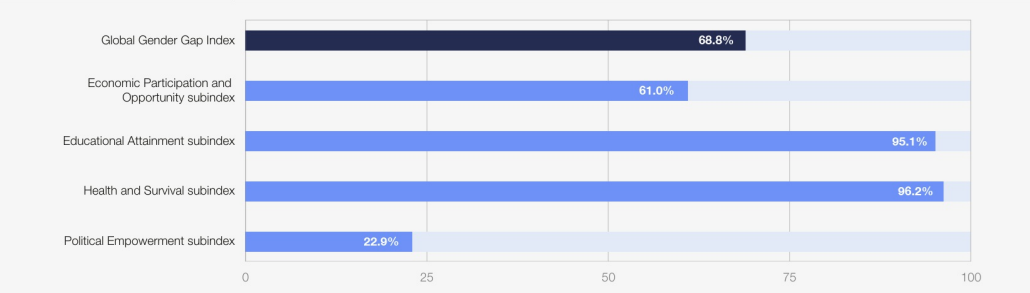

Figure 1: Percentage of the gender gap index (2025).

veloping effective and automated tools for real-time media analysis. To achieve this, Artificial Intelligence (AI) techniques based on Natural Language Processing (NLP) are deployed to carry out regular collection and analysis of articles from news outlets.

In this paper, we describe the various technologies employed in developing the GeMeCo project, emphasizing the importance of utilizing such tools to address and improve the gender gap in media. The article is structured as follows: the next section provides a state-of-the-art review of some of the technologies used, such as co-reference resolution or gender prediction using NLP. Section 3 outlines the methodology followed to develop the GeMeCo project. Then, we describe the details of our evaluation process. Finally, we present the conclusion and suggestions for future work in the last section.

2 *State of the art*

The representation of gender in media has been a subject of extensive investigation. Since 1995 the Global Media Monitoring Project has consistently monitored gender representation in media.

Efforts to automate the process have been successful in Canada for English-language media through the Gender Gap Tracker project (Asr et al., 2021). The Gender Gap Tracker project[7] is an automated system for real-time analysis, that employs NLP technologies to assess the representation of male and female voices within seven major Canadian English-language media outlets. By analysing information gleaned from news articles, the project quantifies the disparity in the proportions of men and women quoted. The project is a collaboration between Informed Opinions,[8] a non-profit organisation committed to increasing women's voices in the media, and the Discourse Processing Lab[9] in conjunction with the Big Data Initiative.[10]

The Gender Gap Tracker project has also been successfully implemented in French under the project called 'Radar de Parité',[11] launched at the end of 2022. This initiative expands the reach of the tracker to French-language media, contributing to a more comprehensive analysis of gender representation across different linguistic contexts, showing that it is possible to transfer the knowledge acquired in the English project (Soumah et al., 2023). Our project proposes adapting the Gender Gap Tracker to the Spanish context to advance gender equality in media analysis. This involves tailoring NLP tools to Spanish, covering reported speech extraction, co-reference resolution, and gender prediction. A review of the state of the art was conducted to ensure the updated system accurately addresses Spanish linguistic characteristics.

2.1 Quotation extraction

The automatic extraction of quotes from news articles has been addressed by several NLP research studies, which have employed a range of techniques, including rule-based methods, machine learning approaches, and hybrid methods. Sagot, Danlos, and Stern (2010) advocated an analysis at the discursive level of French direct quotes that are introduced by a verb in a parenthetical clause. The authors based this analysis on creation of a lexicon of French quotation verbs that was used by a deep linguistic processing chain.

As for applying Machine Learning (ML) to the problem of quotation extraction,

[7] `https://gendergaptracker.informedopinions.org/`

[8] `https://informedopinions.org/`

[9] `https://www.sfu.ca/discourse-lab.html`

[10] `https://www.sfu.ca/big-data.html`

[11] `https://radardeparite.femmesexpertes.ca/`

Fernandes, Motta, and Milidiú (2011) presented a model that was trained on a corpus of Portuguese news articles. The authors decomposed the quotation extraction task into two sub-tasks: identification of quotation boundaries and subsequent association of each identified quotation with its corresponding co-reference set label, which identified the author of the quotation. In this approach, the model is trained to detect the quotation boundaries, reportedly achieving an F1-score of 66.03%.

Despite recent advances in deep learning and the development of powerful pretrained language models, there is still no robust method that would prove to be a reliable state-of-the-art system, especially when working with medium-scale datasets and languages other than English. For this reason, the Gender Gap Tracker and Radar de Parité rely on syntactic and heuristic rules to capture both direct and indirect quotes (Asr et al., 2021; Soumah et al., 2023).

2.2 Co-reference resolution

Co-reference resolution is the process of identifying all expressions within a text that refer to the same entity. It constitutes a critical stage of our analysis, as will be elaborated on in the following sections.

The problem of co-reference resolution has garnered significant attention of the research community for many years. This sustained focus has resulted in the development of numerous tools designed to address this task. One of the first open and flexible systems for co-reference resolution systems was GuiTAR: a General Tool for Anaphora Resolution (Kabadjov, 2007) that was implemented in the MARS knowledge-poor algorithm with sight modifications (Mitkov, 2002). Later, approaches based on unsupervised machine learning were proposed, such as Illinois-Coref, which uses a rich set of features and two types of clustering algorithms (Chang et al., 2011).

Lately, with the advancements in the field of deep learning, neural network-based systems for co-reference resolution were developed, such as NeuralCoref[12] which is built upon pair-wise mention scoring with neural networks that take as input word embeddings and feature representation of each mention surroundings, but only for English. There are ongoing efforts to adapt this system for crosslingual settings and to integrate it into pipelines such as spaCy.[13]

2.3 Gender prediction

Another essential step for every system developed with the objective of performing an analysis of gender disparities in text is gender prediction. Many different studies relied on census databases in order to assign gender to person mentions based on first name, i.e., based on gender assigned at birth. Larivière et al. (2013) review gender disparities in science on name gender assignment performed by accessing several gender name information lists such as the US Census or WikiName, while also making use of a set of language-specific morphological rules. The analysis is, however difficult to reproduce due to non-availability of the full set of names with corresponding gender data. The `genderize.io` API[14] fills the gap in gender prediction software. This software provides access to a regularly updated database of first name and gender correspondences created by harnessing data from major social networks.

We should mention, at this point, that our approach to gender prediction is binary, which we acknowledge is a simplification of the full range of gender expression. We base our analysis on existing databases and on the common association of names with either women or men. Although we include a third category, this includes not just non-binary people, but people for whom we could not reliably identify gender.

Our system extends prior work by adapting quotation and speaker extraction techniques to Spanish, a language for which fewer NLP resources are available. Unlike prior trackers, we account for three types of quotes - direct, indirect (verb-based), and indirect (trigger-based) - as well as interview-based content, which is rarely considered. Additionally, we introduce a hybrid gender prediction module combining official statistics, Wikidata, and online APIs. Finally, our approach integrates daily media scraping and analysis into a real-time dashboard, offering continuous monitoring of gender representation in Spanish digital news media

[12] `https://medium.com/huggingface/state-of-the-art-neural-coref-3302365dcf30`

[13] `https://spacy.io/`

[14] `https://genderize.io/`

3 Methodology

In this section, we provide a detailed overview of the methodology used in our study, focusing on the selection criteria for digital media sources, annotation guidelines, and implementation details. Specifically, we delve into the preprocessing steps, entity recognition, gender prediction techniques, quotation extraction, and coreference resolution methods employed.

3.1 Digital media selection

To study the presence of women in news and their roles as experts or protagonists in the media, we initially considered radio, television, and newspapers. Ultimately, we decided to include six national newspapers in Spain with the largest number of readers according to the latest EGM: *Estudio General de Medios* (General Media Study),[15] although with certain adjustments:

1. Sports newspapers were excluded from our selection due to potential biases.
2. Other media were excluded due to the complexity of web scraping for information extraction.

The selected newspapers were *El País, El Mundo, La Vanguardia, ABC, Público* and *El Periódico*. These are all generalist media with wide coverage of current affairs across the country.

3.2 Annotation guide

Annotation served two goals: it first allowed us to understand and better structure the task. Most importantly, it also helped evaluate the results of the system as we developed it. The annotation focused on identifying named entities that were quoted at least once in the text. Two types of quotations were taken into account for the annotation: direct and indirect quotations. The first type can be identified given that they are enclosed in quotation marks. Here is an example: ***"Nos inquietan los actos administrativos que buscan eludir las vías formales"**, ha argumentado el presidente de los economistas* ("We are concerned about administrative acts that seek to circumvent formal channels," argued the president of the economists.). The second type of quotations are more difficult to detect since they do not have any typographical marks. Most previous research on quotation analysis has only considered direct quotes, while leaving indirect quotes out of the systems, thus skewing the results of any efforts to portray gender representation in media. In order to detect indirect speech, we need to know if the verb that introduces the quote is a declarative verb or verb of speech such as *decir* ('say'), *hablar* ('speak'), *comentar* ('comment'). Another element that can help us detect indirect quotation is a trigger expression, one that introduces a indirect quote without a verb. In Spanish, these include *en su opinión* ('in their opinion') or *según* ('according to'). This is an example: ***En su opinión**, los clientes interesados en la suscripción de vehículos son conscientes de tener en sus manos un automóvil* ('In his opinion, customers interested in vehicle leases are aware of having a car in their hands.').

Thus, reported speech (direct and indirect) was annotated under the label "quotation", and the entities to which the quotations were attributed with the label "reference". The reference is only annotated the first time it is mentioned in the text and all the quotations attributed to it in the text are associated with it by annotation. Annex A provides an example of manual annotation that illustrates how the association between the quotation and the reference is marked. Within the reference tag, we find the attributes: feminine (for references of feminine gender), masculine (for references of masculine gender), and unknown, for those entities whose gender is non-binary or unknown. In this way, we identify the gender of the references so that they can be subsequently counted and compared.

3.2.1 Manual annotation

To create the annotation guide, it was crucial to gather an initial dataset from the six selected newspapers. Crawlers were designed for each media, and news articles were downloaded between June and November 2022. From the available news articles, 20 were randomly selected from each outlet. Out of these 20, only 10 were manually reviewed by two human annotators to ensure they contained at least one quotation. A total of 60 news items (10 per outlet) compose the corpus that we have called GeMeCo.

The development of the annotation guide

[15] https://reporting.aimc.es/index.html#/main/diarios

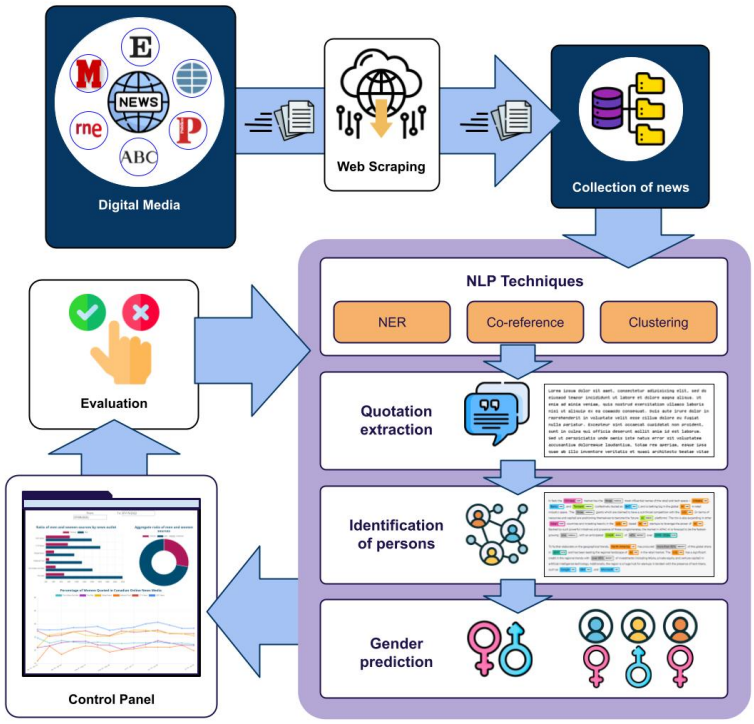

Figure 2: Methodology.

was fully adapted to the experience gained after the annotation of a part of the corpus by six initial annotators: 3 linguists, 2 computer scientists and 1 telecommunications engineer. The Inception platform was used to perform the annotation.[16]

Initially, two news articles were selected and labeled by the six annotators. Challenges encountered during this process were discussed, agreements were calculated, and the annotation guide was adjusted accordingly. This iterative process was repeated twice more, resulting in a total of six news items being used to fine-tune the annotation guide.

To assess the effectiveness of the updated annotation guide, six additional news articles were labeled by the six annotators. This evaluation yielded an agreement rate of approximately 93%.

Two additional linguists were responsible for labeling the entire corpus, consisting of the 60 news items.[17] The agreement rate between the six news items annotated by the initial annotators using the finalised annotation guide and the two new annotators reached 90%.

Inter-annotator agreement was assessed using Cohen's Kappa (κ), applied with the span as the primary annotation unit. The metric was calculated jointly across all experimental subtasks, capturing in a single integrated evaluation the agreement regarding quote boundaries, quotation type, speaker attribution, reference linking, and gender assignment.

The dataset of 60 news articles is considered sufficient to evaluate the module's effectiveness, due to the high volume of words and sentences per article makes manual tagging challenging, as named entities and their quotations often appear in different sentences. Annex B contains statistics on the linguistic data of the corpus.

The results of the corpus analysis reveal a total of 516 citations (237 direct and 279 indirect) linked to 156 different sources as shown in Table 1. In terms of gender distribution, Table 2 illustrates that the profile of the figures is predominantly male (80 cases, equivalent to 51.28% of the occurrence rate), compared to 58 female (37.17%) and 18 unknown. Thus, each news item averages 2.6 references, a presence that once again highlights the gender disparity when comparing

[16] `https://inception-project.github.io/`

[17] The annotation guide is available at `https://github.com/sinai-uja/GeMeCo`

the average for men (1.33) with that for women (0.97).

Quotes	Total	Average per item
Direct	237	3.95
Indirect	279	4.65
Total	516	8.6

Table 1: Corpus Statistics: quotes.

Gender	Total	Average per item
Male	80	1.33
Female	58	0.97
Unknown	18	0.3
Total	156	2.6

Table 2: Corpus Statistics: references per gender.

3.3 Implementation

This subsection provides a comprehensive description of the system developed for analyzing media content. The methodology follows a modular structure, as depicted in Figure 2. The process begins with the collection of news articles from the selected digital media sources via web scraping. Subsequently, the collected data undergoes processing utilizing a range of NLP techniques such as Named Entity Recognition (NER), co-reference resolution, and clustering. These techniques are employed to extract quotes, associate them with their respective speakers, identify their gender, and present the counts on the dashboard.

In the process of finding quotations in the articles, the following information is detected and tagged for each quotation:

- Quote: literal text of the direct or indirect quote.
- Sentence: complete sentence in which the quote appears.
- Speaker: text span within the sentence that identifies the person making the quote (*he, she, the president, the director,* etc.).
- Verb: verb within the sentence that allows the quote to be identified (*said, explained,* etc.).
- Expression: trigger word within the sentence that allows the quote to be identified when the sentence does not have a quotation verb.
- Reference: named entity (first and/or last name) of the speaker. This is the first occurrence of that entity within the article, and may or may not be within the sentence.
- Gender: Gender (woman, man, or unknown) of the named entity.

3.3.1 Pre-processing

The first step performed by our system is text pre-processing, which involves standardizing all quotation marks. This step is intended to reduce typographical errors and improve the precision of direct quote extraction.

3.3.2 Name entity recognition

Following preprocessing, the text of the article undergoes analysis using a Named Entity Recognition (NER) system provided as part of the SpaCy Python package.[18] To achieve high-quality predictions during the NER step, we utilize the `es_core_news_lg` model, which is SpaCy's largest and most accurate CPU-optimized model for Spanish text.[19] We extract entities of type PER (person) from the text, which then undergo a gender identification process.

3.3.3 Gender prediction

Our gender prediction module is based on names and relies on an internal database initially populated with data from the Spanish National Institute of Statistics. This database provides a list of first names associated with their respective gender.[20] However, the database is insufficient for cases where a first name is ambiguous (e.g., Alex) or when the entity contains more than just a first name.

In cases where entities are not found in our internal database, the system automatically queries Wikidata[21] to retrieve the 'sex or gender' attribute associated with the entity. If no gender information is available

[18] https://spacy.io/

[19] https://spacy.io/models/es#es_core_news_lg

[20] https://ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177009&menu=resultados&idp=1254734710990

[21] https://www.wikidata.org/

from Wikidata, the system then calls upon Genderize.[22]

All results obtained from the calls to Wikidata and Genderize are stored in our internal database to minimize the need for future external service calls and to improve efficiency.

3.3.4 Extraction of quotes

In order to achieve the previously mentioned objective of this study, accurately counting quotes plays a key role. For this purpose, we developed a ruled-based system whereby we search all instances of quotes within the article. In this module, we distinguish between direct and indirect quotes. Specifically, within indirect quotes, we further differentiate between those with a verb and those with a trigger. Annex C provides examples of quotes types that follow this rule-based system.

For direct quotes, we search for opening quotation marks. Once identified, we locate the corresponding closing quotation mark to extract the quoted information. Additionally, we capture the sentences surrounding the quote to identify the verb and speaker. In cases where the sentence lacks a verb, we search for the expression containing the subject or speaker of the quote. Figure 4 in Annex D illustrates the flowchart of the direct quotations extraction process.

In the case of indirect quotations, what we are seeking is the verb or trigger that usually goes with this type of quotation. Once obtained, the process is the same as in the case of direct quotations, but applying different linguistic rules. Figures 5 and 6 in Annex D show the flowcharts describing the extraction of indirect quotes by detecting the verb or detecting the trigger.

After obtaining these data, we apply the co-reference explained in the following section (3.3.5) and assign each quote to a reference and its gender.

Some articles follow an interview format where responses (R that stands from the Spanish word 'Respuesta') and questions (P that stands from the Spanish word 'Pregunta') are distinguished. All articles are previously processed by a function that extracts their quotations in case they have this format. Therefore, the regular expressions "R." and "P." are used, the first of these being considered as the beginning of the quotation and the second as the end. Once found, the content between both expressions is extracted as a quotation. In this type of articles, it can be observed that they usually begin by introducing the interviewee. From this, each quote is assigned the first entity found in the text along with its gender. Figures 7 in Annex D shows the flowchart describing the process of extracting quotes from the interviews.

Finally, the quote extraction pipeline varies according to the document type: for interviews, it performs a dedicated interview quotes extraction; otherwise, it executes a sequential extraction of direct quotes, followed by indirect quotes using both verbs and expressions.

3.3.5 Co-reference resolution

For the co-reference resolution we developed a module in which we use the crosslingual co-reference model.[23] This model provides clusters containing the different occurrences of the entities in the text that the model detects. The model also provides a dictionary containing the representative of each of these clusters and their occurrences.

Initially, leveraging the entities previously detected by spaCy, we construct a cluster for each entity that includes its occurrences in the text, gender information, and other characteristic attributes mentioned in the text associated with that entity. Subsequently, we merge this cluster with the cluster provided by the model, yielding a consolidated cluster containing an identifier, associated entities, occurrences, and gender information. This cluster will be used to search the speakers obtained in the extraction of quotations and assign each quotation to an entity and a gender.

4 Evaluation

4.1 Quote evaluation

To evaluate quotation extraction, we compare our system's output with annotations made by human annotators. This comparison involves analyzing the span, in characters, of both the annotated and extracted quotes. We define a match between these spans as follows, where S_a represents the span of the annotated quote and S_e represents the span of the extracted quote:

[22] https://genderize.io/

[23] https://github.com/davidberenstein1957/crosslingual-coreference

Subcorpus (threshold)	P	R	F1
Corpus 1 (0.3)	0.533	0.802	0.640
Corpus 1 (0.8)	0.508	0.765	0.611
Corpus 2 (0.3)	0.435	0.528	0.477
Corpus 2 (0.8)	0.333	0.404	0.365
Corpus 3 (0.3)	0.603	0.864	0.711
Corpus 3 (0.8)	0.527	0.754	0.620
Corpus 4 (0.3)	0.571	0.645	0.606
Corpus 4 (0.8)	0.467	0.527	0.495
Corpus 5 (0.3)	0.395	0.634	0.486
Corpus 5 (0.8)	0.316	0.507	0.389
TOTAL 0.3	**0.507**	**0.695**	**0.584**
TOTAL 0.8	**0.430**	**0.591**	**0.496**

Table 3: Quote evaluation.

$$Score = \frac{len(S_a \cap S_e)}{len(S_a)} \quad (1)$$

We use two thresholds (0.3 and 0.8) to assess the degree of overlap between an annotated quote with a predicted one. The quote is counted as predicted as long as the score obtained exceeds the established threshold. For example, with a score of 0.6 the quote would match the 0.3 threshold but not the 0.8.

In order to perform the experimentation and to avoid over-fitting the algorithms, the collection of news articles was divided into 5 sub-corpora of 12 articles each. Table 3 shows the results obtained. The evaluation shows that, although the system can be improved, it detects the majority of the quotes (average recall of 0.7 for the 0.3 threshold), with improvements needed in the precision.

4.2 Unique references

In this step, we check the reference detection task. Instead of checking the reference (full name) of each speaker associated with each citation, we count the distinct references detected in the whole article. The main objective is to extract the unique references that have been taken into consideration in each article, and ensure high precision and recall. For this reason, we check who has been quoted both in the gold standard and in what the system predicts, without taking into consideration the number of quotes by each speaker.

A unique reference is considered to be correct when, after lowercasing, the Levenshtein distance is less than 2 between the system's speaker and the annotation, i.e. they are practically the same. This is counted as true positive (*Correct Reference*). Also, the total number of different references predicted and annotated references are counted. With this data, we calculate the precision, recall and F1 metrics.

We calculated these metrics for each of the five subcorpora and the total corpus, obtaining the results shown in Table 4. As we can see, we achieve fairly good precision. The main problems observed come from the co-reference resolution. In many cases, the model was not able to match a speaker to the initial reference. In some other cases, the speaker could not be detected and a reference could not be established.

	Precision	Recall	F1
Corpus 1	0.611	0.440	0.511
Corpus 2	0.923	0.363	0.521
Corpus 3	0.739	0.369	0.492
Corpus 4	0.857	0.352	0.500
Corpus 5	0.785	0.458	0.579
TOTAL	**0.776**	**0.395**	**0.524**

Table 4: Speaker metrics.

In the case of recall, one of the main problems was in linking the speaker with the correct reference. In most cases, the speaker was linked to the wrong reference. This is usually due to the coreference clusters returned by the model. Some of these clusters contain more than one entity, but only one is the representative entity. Therefore, in the case where we have detected as a speaker some entity that is not a representative, but is in the cluster, the reference will be incorrectly assigned.

The other problem is in cases where that entity has only one quote in the article and the system was not able to identify it.

4.3 Gender prediction

To find out the capacity of our system to predict gender, we carried out two different evaluations. The first one is more specific to the gender prediction module. For this, we calculated the precision, recall and F1 in order to see the rate of success of this part of our system. The gender is predicted for each speaker detected. The precision and recall values in Section 4.2 (Speaker results) show that we have incorrectly predicted speakers. However, these speakers were automatically

gender tagged and we cannot check if gender prediction is correct because they are not gender annotated. For this reason, to obtain the precision, recall and F1 metrics, we calculated the number of references that have the same gender (*True Positives*), the number of predicted references of a gender that are incorrect (*False Positives*) and the number of references noted that were not correctly predicted (*False Negatives*) from the intersection of the predicted and annotated references for both male and female. The results obtained for each gender can be seen in Table 5.

	Precision	**Recall**	**F1**
Men	1.000	0.927	0.962
Women	1.000	0.937	0.968

Table 5: Gender metrics.

For the second evaluation, we adopted a broader approach, aiming to determine the ratio of men and women predicted by our system compared to the ratio annotated by human annotators. As we can see in Table 6, the ratios are quite comparable, with a slight bias towards predicting more women than men, which we would like to explore in future work.

	Annotated	**Predicted**
Men	0.64	0.62
Women	0.36	0.38

Table 6: Gender ratio.

5 The GeMeCo dashboard

All the modules described in the previous sections have been integrated to generate a web dashboard where the results of our research can be displayed. The GeMeCo dashboard enables visualization of the proportion of women and men quoted in news from Spain's most influential online media through various graphs. Data is collected daily and presented on the GeMeCo interactive website,[24] where users can input a time interval to automatically generate several graphs with the corresponding data. In case no specific date is provided, the default display shows results from seven days prior to the current date, counting up to three days before the current date. The graphs reveal that, on average, about 30% of the sources are female, while 70% are male. These percentages persist even when changing time intervals, reflecting a consistent pattern over the monitored period.

Figures 8 and 9 in Annex E provide an example of how the data are presented within the time interval from March 20, 2024 to April 20, 2024, for the six selected online newspapers.

Figure 8 illustrates the number of male and female sources found in each of the 6 newspapers on one side, while on the right-hand side, it shows the overall proportion of men and women quoted. On the other hand, Figure 9 focuses on the percentage of women quoted in each media outlet.

6 Conclusion

While gender equality is often discussed as an achieved milestone, the reality remains that women are significantly underrepresented across most sectors of global society compared to men. The GeMeCo project addresses this persistent gender gap by focusing specifically on the Spanish digital media landscape. By leveraging advanced AI techniques based on NLP, such as Named Entity Recognition (NER) and co-reference resolution, we developed a specialized tool to monitor and quantify the presence of men and women in the news.

Through a dedicated interface, the tool analyzes both direct and indirect quotations from six major Spanish digital newspapers, highlighting a stark disparity: women are sourced approximately 30% of the time, while men account for the remaining 70%. As future work, we intend to maintain regular monitoring and system calibration to ensure data accuracy over time. For this purpose, we maintain a database of names and their gender attribution, which we update with new entries. This stored information serves as a longitudinal indicator to assess the evolution of the gender gap. Furthermore, we aim to scale the GeMeCo methodology to include a broader range of media outlets and even the official proceedings of regional government parliaments across Spain.

Acknowledgements

This work is funded by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia -

[24] https://gemeco.ujaen.es/

Funded by EU – NextGenerationEU within the framework of the project Desarrollo Modelos ALIA. This work has also been partially supported by Project CONSENSO (PID2021-122263OB-C21) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR, Project ROMANET (CERV-2024-CHAR-LITI-101215052), funded by the European Union under the Citizens, Equality, Rights and Values programme, Project HEART-NLP-UJA (PID2024-156263OB-C21) and Project VERITAS-H (AIA2025-163322-C64) funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU, Project GALENO-IA (DGP_PIDI_2024_00852) funded by the European Union under the Smart Specialization Strategy Program for Sustainability in Andalusia (S4Andalucía) 2021–2027.

References

[Asr et al.2021] Asr, F. T., M. Mazraeh, A. Lopes, V. Gautam, J. Gonzales, P. Rao, and M. Taboada. 2021. The gender gap tracker: Using natural language processing to measure gender bias in media. *PloS one*, 16(1):e0245533.

[Berenbaum2019] Berenbaum, M. R. 2019. Speaking of gender bias. *Proceedings of the National Academy of Sciences*, 116(17):8086–8088.

[Chang et al.2011] Chang, K.-W., R. Samdani, A. Rozovskaya, N. Rizzolo, M. Sammons, and D. Roth. 2011. Inference protocols for coreference resolution. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 40–44.

[Fernandes, Motta, and Milidiú2011] Fernandes, W. P. D., E. Motta, and R. L. Milidiú. 2011. Quotation extraction for portuguese. In *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology*, pages 204–208.

[Gordon2002] Gordon, L. 2002. *The moral property of women: A history of birth control politics in America*. University of Illinois press.

[Kabadjov2007] Kabadjov, M. A. 2007. *A comprehensive evaluation of anaphora resolution and discourse-new classification*. Ph.D. thesis, Citeseer.

[Kassova2020] Kassova, L. 2020. The Missing Perspectives of Women in News: A report on women's under-representation in news media; on their continual marginalization in news coverage and on the underreported issue of gender inequality. Report, AKAS Consulting.

[Kassova2023] Kassova, L. 2023. From Outrage to Opportunity: How to include the missing perspectives of women of all colors in news leadership and coverage. Report, AKAS Consulting.

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

[Mitkov2002] Mitkov, R. 2002. *Anaphora Resolution*. Routledge.

[Sagot, Danlos, and Stern2010] Sagot, B., L. Danlos, and R. Stern. 2010. A lexicon of french quotation verbs for automatic quotation extraction. In *7th international conference on Language Resources and Evaluation-LREC 2010*, pages 294–299.

[Schuster2017] Schuster, J. 2017. Why the personal remained political: comparing second and third wave perspectives on everyday feminism. *Social Movement Studies*, 16(6):647–659.

[Soumah et al.2023] Soumah, V.-G., P. Rao, P. Eibl, and M. Taboada. 2023. Radar de Parité: An NLP system to measure gender representation in French news stories. In *Proceedings of the 36th Canadian Conference in Artificial Intelligence*, pages 1–12, Montréal.

A Annex: Annotation platform

Figure 3 is part of a text annotation from the corpus using the Inception platform.

B Annex: Information of corpus

Table 7 shows the average number of sentences (#Sentences), the average number of words (#Words) and the number of words per sentence (#(Wor/Sen) per newspaper article.

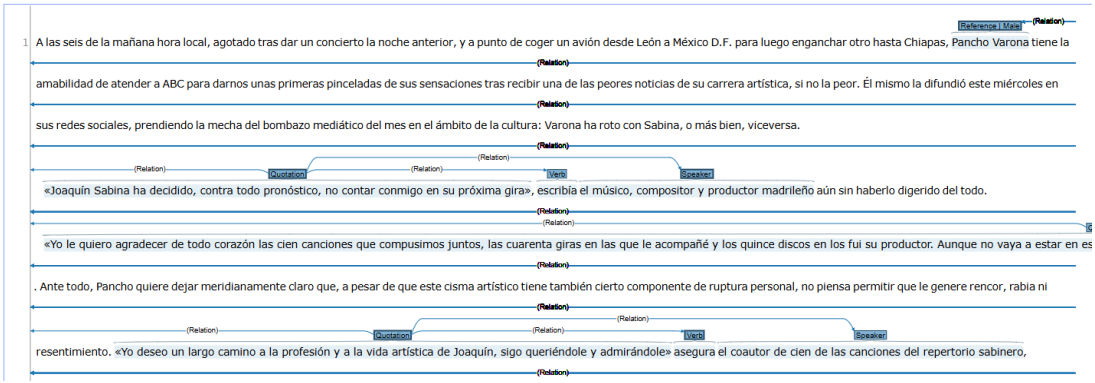

Figure 3: Annotated text from the GeMeCo corpus using the Inception platform.

Newspaper	#Sentences	#Words	#Word/Sen
El País	60.4	1265.5	25.64
El Mundo	17.4	535.8	30.60
La Vanguardia	29.9	791.4	29.11
ABC	38.7	838.6	27.6
Público	38	1220.7	40.96
El Periódico	28	732.2	27.24

Table 7: Linguistic features per newspaper (average per article).

C Annex: Examples of types of quotes

This Annex shows several examples of quotations.

- Example of direct quote in bold:

  *Pancho comentó a esta periodista "**No siento que Sabina me haya hecho un feo**".* [Pancho told this reporter, "**I don't feel like Sabina snubbed me**".]

- Example of indirect quote with verb in bold:

  *Pancho comentó a esta periodista que **no sentía que Sabina le hubiese hecho un feo**.* [Pancho told this reporter that he **didn't feel like Sabina had snubbed him**.]

- Example of indirect quote with trigger in bold:

  *Según la periodista, Pancho **no sentía que Sabina le hubiese hecho un feo**.* [According to the reporter, Pancho **didn't feel like Sabina had snubbed him**.]

D Annex: Flowcharts

This Annex contains the flowcharts used for quote extraction.

E Annex: Dashboard

This Annex provides an example of how the data are presented in the GeMeCo dashboard.

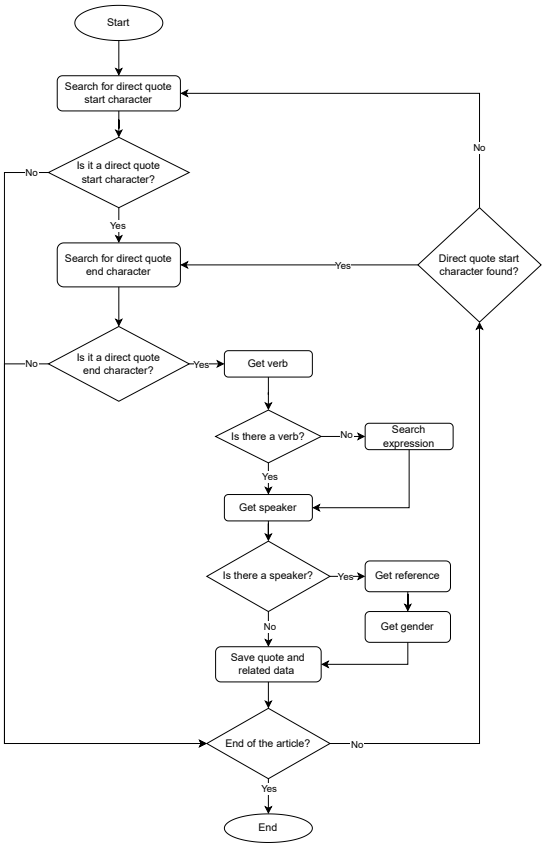

Figure 4: Direct quote extraction.

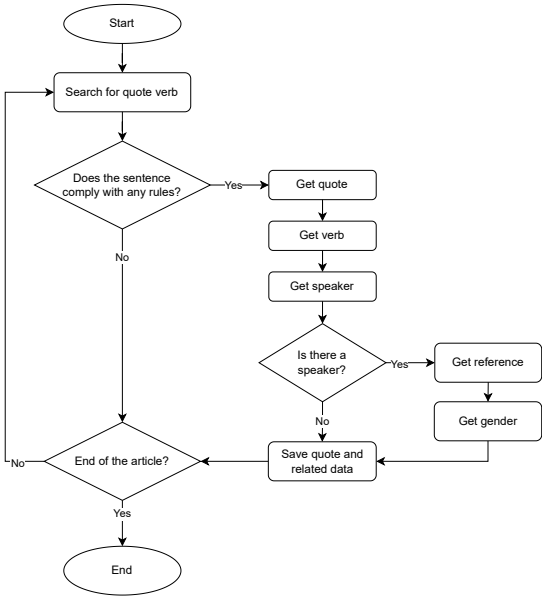

Figure 5: Indirect quotes extraction with verbs.

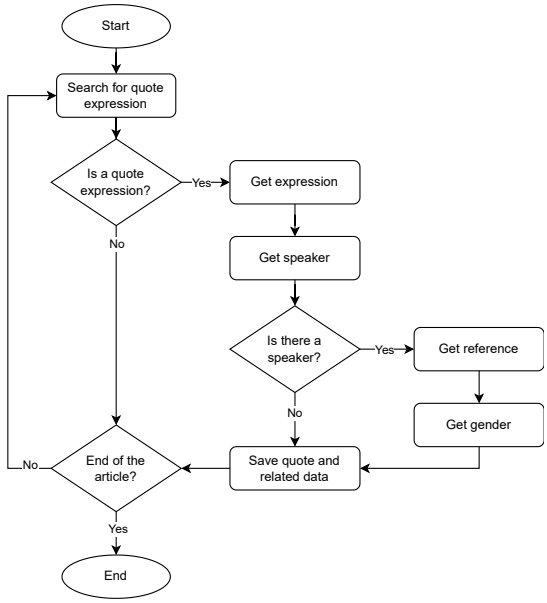

Figure 6: Indirect quote extraction with trigger.

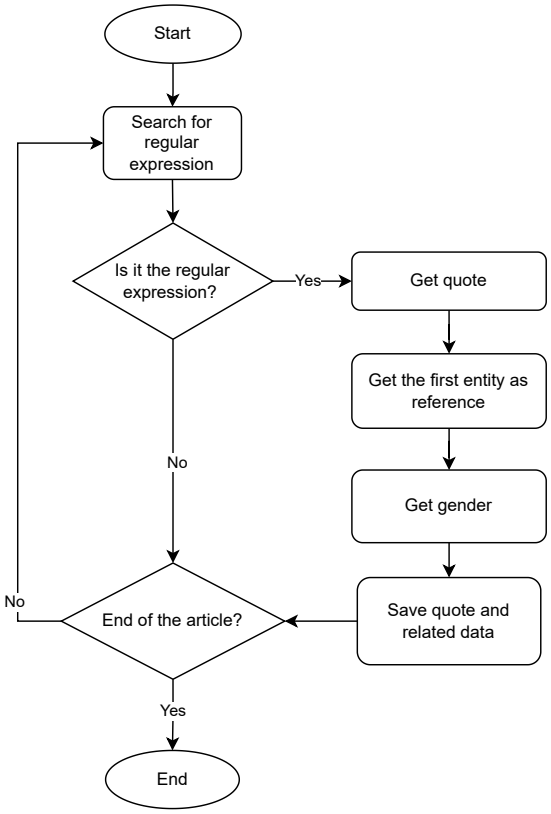

Figure 7: Interview quotes extraction.

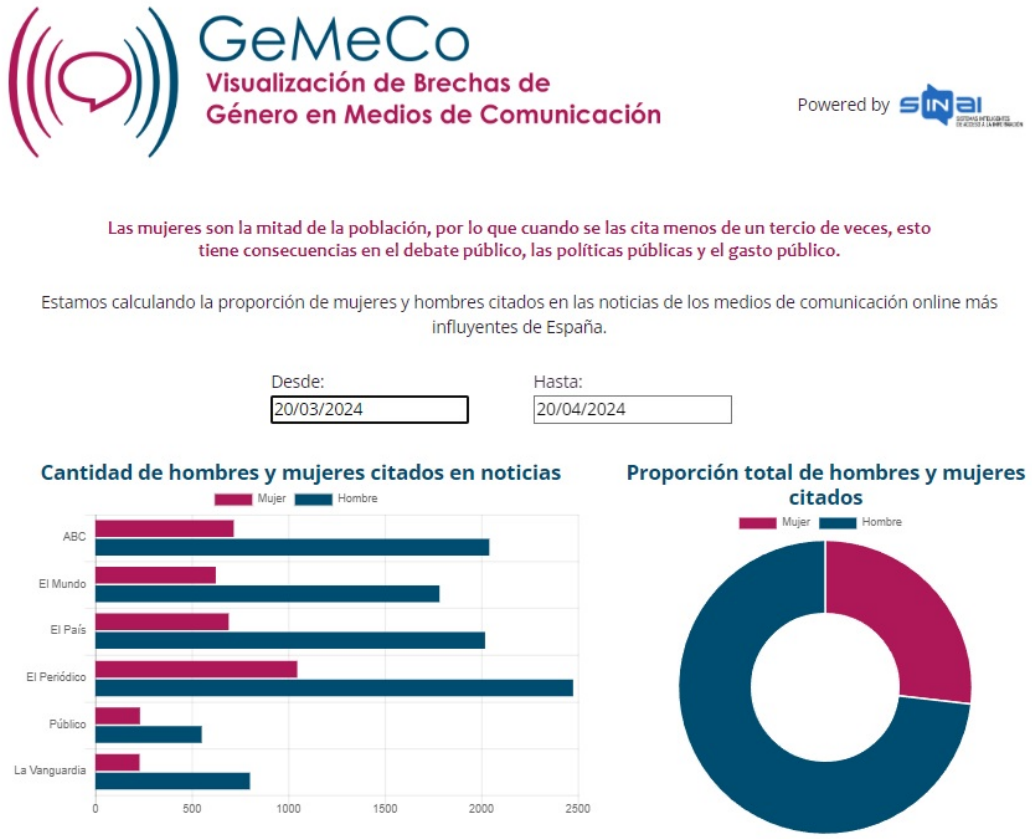

Figure 8: Percentages of quotations by men and women.

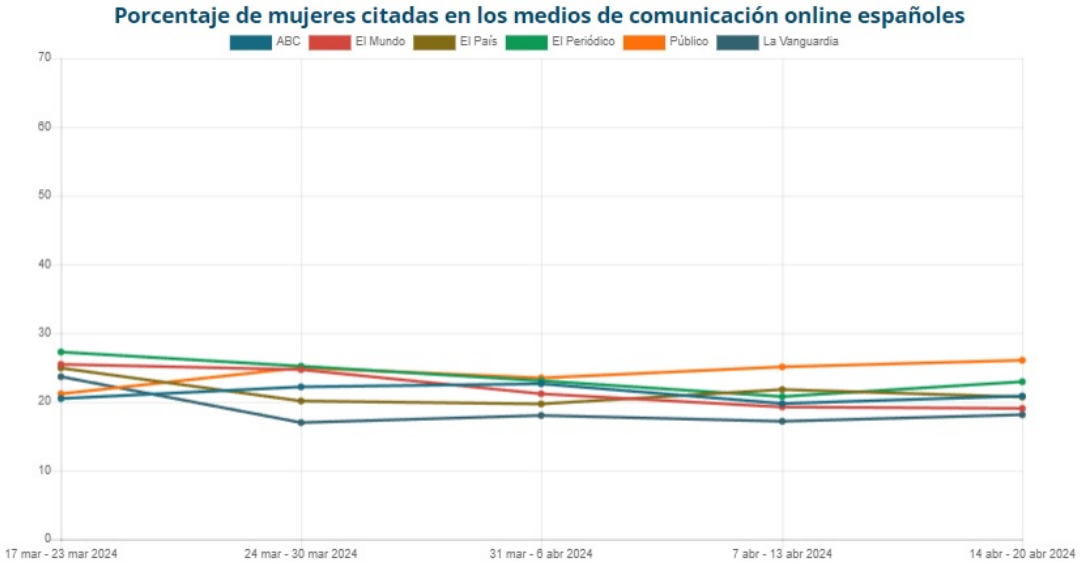

Figure 9: Percentage of women quoted in each digital media.