

Detecció i mesura de biaixos de gènere en els models *deep learning* en català.

UOC

Nura Mangado Gómez

Grau de Ciència de Dades aplicades.
Processament del llenguatge natural.

Tutor

David Puentes Jurado

Professor responsable de l'assignatura

David Merino Arranz

Data d'entrega

01/2023



Aquesta obra està subjecta a una llicència de
Reconeixement-NoComercial-SenseObraDeivada
[3.0 España de Creative Commons](#)

FICHA DEL TRABAJO FINAL

Títol del treball:	Detecció i mesura de biaixos de gènere en els models del llenguatge <i>deep learning</i> en català.
Nom de l'autor:	Nura Mangado Gómez
Nom del consultor:	David Puentes Jurado
Nom del PRA:	David Merino Arranz
Data d'entrega (mm/aaaa):	01/2023
Titulació o programa:	Grau de Ciència de Dades Aplicades
Àrea del Treball Final:	Processament del llenguatge natural
Idioma del treball:	Català
Paraules clau	Processament del llenguatge natural, models <i>deep learning</i> , biaixos

Resum del Treball

En els darrers anys hi ha hagut grans avanços en el món de la ciència de dades i la intel·ligència artificial i, concretament, en el camp del processament del llenguatge natural (PLN). En aquest sentit, el desenvolupament dels models del llenguatge *deep learning* ha suposat una millora important en moltes aplicacions de NLP, així com la seva implementació a llengües no majoritàries com és el català.

Paral·lelament, una part de la societat s'està movent per deixar enrere les desigualtats socials, malgrat que aquestes han deixat una empremta en la cultura i en el llenguatge, que alhora es retroalimenten. En el cas del llenguatge, hi ha una sèrie de biaixos que afavoreixen estereotips afectant la manera com concebem el món. Aquests biaixos es veuen doncs reflectits en els models del llenguatge, sent essencial detectar-los i erradicar-los. Aquest treball té com a objectiu mostrar com alguns biaixos de gènere inherents a la societat s'han propagat als models del llenguatge en català.

La metodologia seguida és una adaptació dels mètodes de Kurita i Bartl et al., al model català (BERTA), on es mostra l'associació entre professions i gènere a través de la creació de frases emmascarades. Tot i que s'han detectat diferències en la resposta del model en funció del gènere, la metodologia no permet mesurar de forma clara aquests biaixos com ho fa amb l'anglès, a causa de les marques de gènere que inclou el llenguatge català, deixant per a futurs estudis l'exploració de nous mètodes.

Abstract

In recent years, there have been great advances in the world of data science and artificial intelligence, and specifically in the field of natural language processing (NLP). In this sense, the development of deep learning-based language models has increased many NLP applications performance, as well as its implementation in non-majority languages such as Catalan.

At the same time, a part of the society is working to leave social inequalities behind, even though those inequalities have already left a mark on culture and language, mutually reinforcing each other. In language, there are a number of biases that create stereotypes affecting the way we conceive the world. These biases are therefore reflected in language models, and it is essential to detect and eradicate them. This work aims to show how some gender biases inherent in society have spread to language models in Catalan.

The methodology followed is an adaptation to the Catalan model (BERTA) of Kurita and Bartl et al.'s methods, where the association between professions and gender is analysed through masked sentences. Although differences have been detected in the response of the model based on gender, the methodology does not allow to clearly measure these biases as it does with English, due to the gender markers included in the Catalan language, leaving for future studies the exploration of new methods.

Gràcies a Maria i família, per tot el suport incondicional.

The screenshot shows the Google Translate website interface. At the top, there is a navigation bar with three horizontal lines and the text "Google Translate". Below this are two tabs: "Text" (selected) and "Websites". The language pair is set to "ENGLISH" to "CATALAN". The main area contains the following text:

A father and his son are in a car accident. The father dies at the scene and the son is rushed to the hospital. At the hospital the surgeon looks at the boy and says "I can't operate on this boy, he is my son." How can this be?

Below this, there is a blue highlighted section containing the same text in Catalan:

Un pare i el seu fill tenen un accident de trànsit. El pare mor al lloc dels fets i el fill és traslladat d'urgència a l'hospital. A l'hospital el cirurgià mira el nen i li diu: "No puc operar aquest noi, és el meu fill". Com pot ser això?

At the bottom of the interface, there are several icons: a microphone, a speaker, a refresh symbol, a star, a magnifying glass, and a left arrow.

II·lustració 1 - Traducció endevinallla cirurgià (Google Translate). Consulta 21-12-2022.

Índex

1. Introducció	6
1.1. Context i justificació del Treball.....	6
1.2. Objectius del Treball.....	7
1.3. Impacte en sostenibilitat, ètica-social i de diversitat.....	8
1.4. Enfocament i mètode seguit.....	10
1.5. Planificació del treball.....	12
1.5.1 Anàlisi i gestió de riscos.....	13
2. Processament del llenguatge natural.....	15
2.1. Representació de paraules	16
2.2. Representació de paraules segons el context, els <i>Word embeddings</i>	16
2.3. <i>Word embeddings</i> i <i>deep learning</i>	19
2.3.1 <i>BERTa, el model català</i>	20
2.4. Ètica, intel·ligència artificial i PLN	23
2.4.1. <i>Mesura de biaixos de gènere en els models del llenguatge</i>	24
2.5. Funcionament dels models tipus BERT	26
2.5.1. <i>Tokenització</i>	26
2.5.2. <i>Funcionament dels models BERT i RoBERTa i la tasca fill-mask</i>	28
3. Materials i mètodes	31
3.1 Recol·lecció de dades de professions.....	31
3.2 Neteja i preparació de dades de professions	32
3.3 Adaptació del mètode de Bartl	35
3.3.1 <i>El mètode de Bartl</i>	35
3.3.2 <i>Plantilles tipus per crear les frases</i>	35
3.3.3 <i>Termes relatius a persones</i>	36
3.4 Càlcul d'associació entre professions i gènere pel model BERTa.	38
4. Resultats.....	45
4.1 Professions amb marca gramatical de gènere	45
4.1.1. <i>Frases amb gènere emmascarat</i>	45
4.1.2. <i>Frases amb gènere desemmascarat - Professions amb marca de gènere gramatical</i>	50
4.1.3. <i>Associació - Professions amb marca de gènere</i>	50
4.2 Professions sense marca de gènere grammatical (49 professions)	52
5. Conclusions i treballs futurs	59
5.1 Lliçons apreses	Error! Bookmark not defined.

5.2 Objectius assolits	60
5.3 Competència de “Compromís ètic i global”.....	60
5.4 Línies de futur	60
6. Glossari	62
7. Bibliografia	64
8. Annexos.....	68

Llista de figures

Figura 1 - Funcionament BERTa, tasca <i>fill-mask</i>	30
Figura 2 - <i>Wikidata Query Service</i>	Error! Bookmark not defined.
Figura 3 - Distribució percentatge dona/home per professions Wikidata.	34
Figura 4 - Comparació de probabilitats en context sense gènere (paraules masculines/femenines diferents).	45
Figura 5 - Distribució de probabilitats en context sense gènere.....	46
Figura 6 - Distribució de probabilitats en context sense gènere (professions "femenines").....	47
Figura 7 - Probabilitats ballarina/ballarí (context sense gènere) agrupat per plantilla.....	48
Figura 8 - Probabilitats ballarina/ballarí (context sense gènere) agrupat per plantilla (escala log.)	Error! Bookmark not defined.

Llista d'il·lustracions

Il·lustració 1 - Traducció endevinalla cirurgià (Google Translate)	1
Il·lustració 2 - Exemple ChatGPT	6
Il·lustració 3 - Traducció endevinalla cirurgià (Aina MT) Error! Bookmark not defined.	
Il·lustració 4 - Consulta SPARQL a <i>Wikidata Query Service</i> Error! Bookmark not defined.	

Llista de taules

Taula 1 - Objectius específics.....	8
Taula 2 - Exemple relació paraules - tokens (BERTa)	28
Taula 3 - Tokenització per al model BERTa de la frase “Ella és <mask>.”	29
Taula 4 - Valors retornats per el model BERTa per la posició 3 i la frase “Ella és <mask>.”	29
Taula 5 - Professions Wikidata agrupades en actor/actriu.	33
Taula 6 - Etapes neteja de dades de professions <i>Wikidata</i>	34
Taula 7 - Plantilles de frases Bartl et al. (2020).. Error! Bookmark not defined.	
Taula 8 - Plantilles de frases utilitzades en l'estudi	36
Taula 9 - Termes de persones utilitzats per Bartl et al. (2020)[1].....	35

Taula 10 - Noms de persones més comuns al model BERTa.....	37
Taula 11 - Termes de persones utilitzats.....	37
Taula 12 – Etapes del càlcul d'associació professions-gènere (professions amb marcat de gènere).	38
Taula 13 - Etapes del càlcul d'associació professions-gènere (Bartl et al., 2020)	39
Taula 14 - Exemple de frase persona - professió.....	39
Taula 15 - Exemple de frase amb la paraula objectiu (professió) emmascarat.	40
Taula 16 - Exemple de frase amb la paraula objectiu i gènere emmascarat....	40
Taula 17 - Exemple de probabilitat actor/actriu en un context amb gènere.	41
Taula 18 - Exemple de probabilitat actor/actriu en un context sense gènere..	41
Taula 19 - Exemple d'associació actor/actriu - persona.	42
Taula 20 - Exemple de frase amb la paraula objectiu (gènere) emmascarada.	42
Taula 21 - Exemple de frase amb la paraula objectiu i professió emmascarades.	42
Taula 22 - Exemple de probabilitat pare/mare en un context amb l'atribut professió.	43
Taula 23 - Exemple de probabilitat actor/actriu en un context sense gènere..	43
Taula 24 - Exemple d'associació actor/actriu - persona.	44
Taula 25 – Mitjana de probabilitats en context sense gènere (majoria femení)	47
Taula 26 - Annex A - Professions no reconegudes pel model BERTa.	69

1. Introducció

1.1. Context i justificació del Treball

La intel·ligència artificial (IA) està present en el nostre dia a dia en multitud d'aplicacions, des de xatbots o assistents veu, detectors de reconeixement facial o en sistemes de presa de decisió. Entrat l'any 2023, ens trobem a les portes d'una nova revolució industrial on la IA està assolint uns nivells de rapidesa i precisió en les respostes mai vista fins al moment. Una mostra d'això es la recent aparició de ChatGPT¹ amb capacitat per programar i escriure, en què la qualitat de les respostes situa aquesta tecnologia en una realitat que fins al moment estava reservada per a tasques molt concretes.

The screenshot shows a user input "NU write an introduction about gender bias in NLP" followed by a generated response from ChatGPT. The response discusses gender bias in NLP, mentioning how machine learning models can perpetuate or amplify harmful biases found in society. It highlights various ways this can happen, such as through training data, model design, and output. The response concludes with a note on increasing awareness and efforts to address gender bias in NLP.

NU write an introduction about gender bias in NLP

Gender bias in natural language processing (NLP) refers to the ways in which machine learning models that process and analyze language data can perpetuate or amplify harmful gender biases that exist in society. These biases can manifest in a variety of ways, such as through the language used in training data, the design and development of models and algorithms, and the output of NLP systems. This can lead to discriminatory outcomes and a lack of fairness in the way that NLP systems are used. In recent years, there has been increasing awareness and efforts to address gender bias in NLP, including through the development of fairer and more inclusive models and the use of diverse and representative training data.

II·lustració 2 - Exemple ChatGPT

La proliferació de les funcionalitats de les noves tecnologies ha comportat grans avanços, però també ha accentuat algunes problemàtiques com són les desigualtats i els biaixos socials i de gènere, que queden reflectides en aplicacions basades en intel·ligència artificial. Aquests sistemes d'IA, no només copien aquests biaixos sinó que sovint els amplifiquen [1]. Malgrat tot, la comunitat científica, empentada per una part de la societat activa i preocupada per les desigualtats socials i l'àtica de la intel·ligència artificial, està estudiant aquesta problemàtica i actuant per corregir-la [2].

Paral·lelament, en l'àmbit català, el Departament de la Vicepresidència i de Polítiques Digitals i Territori del Govern de Catalunya, amb la col·laboració del Barcelona Supercomputing Center (BSC), ha impulsat el projecte AINA², que té com a objectiu generar recursos i models de la llengua catalana per a ser utilitzats en aplicacions basades en intel·ligència artificial i processament del llenguatge natural PLN.

¹ <https://chat.openai.com/>

² <https://politiquesdigitals.gencat.cat/ca/economia/catalonia-ai/aina/>

Aquesta problemàtica només es podrà resoldre si la comunitat científica, les empreses tecnològiques que desenvolupen aquests models, els governs i tota la societat s'impliquen per aconseguir-ho. Aquest treball pretén fer visible una part d'aquesta problemàtica mitjançant la **detecció de biaixos de gènere en els models de deep learning en català**, com a primer pas necessari per determinar si aquests models són **aptes o no per ser aplicats**. Addicionalment, remarcar la importància de fer més i més rigorosos controls ètics als models d'IA i no limitar-se a les mètriques d'avaluació clàssiques.

Aquest treball té un impacte potencial en diversos àmbits. Per una banda, en l'àmbit **acadèmic**, ja que, donat que la majoria d'estudis de biaixos de gènere són pel llenguatge anglès, aquest treball pot ser **d'utilitat per estudiar llengües no majoritàries com és el cas del català**. Al mateix temps, servirà per mostrar la validesa de les tècniques de detecció de biaixos estudiades i aprovades en la llengua anglesa. En segon lloc, aquest treball té un **interès social**, ja que pretén determinar si els models, desenvolupats i publicats en el marc del projecte AINA, estan **lliures de biaixos i són, per tant, aptes per a aplicar-se**.

Finalment, aquest treball també té un **interès econòmic** o de negoci, ja que aquests models són utilitzats per **desplegar aplicacions d'IA** i, en conseqüència, aquelles empreses que els estiguin utilitzant i que vulguin usar models lliures de biaixos de gènere, poden tenir interès en els resultats d'aquest treball.

Així, aquest treball es nodreix de diverses temàtiques com són l'ètica en la IA mitjançant el tractament dels biaixos de gènere, l'estudi dels models del llenguatge emprats en el processament del llenguatge natural i l'estudi dels models en una llengua no majoritària com és el català. Amb la combinació d'aquests àrees d'estudi, es presenta aquest treball que vol estudiar l'existència de biaixos de gènere en els models del llenguatge català.

1.2. Objectius del Treball

Objectiu principal

L'objectiu principal del treball és l'estudi de l'existència i quantificació dels biaixos de gènere en els models del llenguatge català creats pel Barcelona Supercomputing Center (BSC) en el marc del projecte AINA.

Objectius generals

1. Estudi de l'estat de l'art de les tècniques existents de detecció i mesura de biaixos en els models del llenguatge.
2. Adaptació de les tècniques estudiades als models del llenguatge en català seleccionats.

Objectius específics

D'acord amb els objectius generals, s'han definit els següents objectius específics que es mostren a la **taula 1**.

1. Estudi de l'estat de l'art de les tècniques existents de detecció i mesura de biaixos en els models del llenguatge.

- 1.1. Estudiar les diferents tipologies de models del llenguatge.

Entendre el funcionament dels models del llenguatge i adquirir els coneixements tècnics de les eines necessàries per a la utilització dels models deep learning.

- 1.3. Recopilar la documentació acadèmica sobre l'estudi de biaixos existents en els models del llenguatge.

2. Adaptació de les tècniques estudiades als models del llenguatge en català seleccionats.

- 2.1 Estudiar les característiques tècniques, la tipologia i el funcionament dels models seleccionats.

- 2.2 Adaptar les tècniques recopilades als models del llenguatge seleccionats.

- 2.3 Elaborar un informe amb els resultats i conclusions.

- 2.4 Proposar mesures per mitigar els biaixos.

Taula 1 – Objectius generals i específics.

1.3. Impacte en sostenibilitat, ètica-social i de diversitat

D'acord amb el compromís de la Universitat Oberta de Catalunya (UOC)³ de contribuir a l'Agenda 2030 amb els Objectius de Desenvolupament Sostenible (ODS) de l'ONU⁴, en aquest apartat es descriu l'alineació d'aquest treball amb les tres dimensions de la competència transversal UOC “Compromís ètic i global”:

- Sostenibilitat
- Comportament ètic i responsabilitat social (RS)
- Diversitat (gènere entre d'altres) i drets humans”)

³ <https://www.uoc.edu/portal/es/compromis-social/index.html>

⁴ <https://www.un.org/sustainabledevelopment/>

Sustainable Development Goals



II·lustració 3 - Objectius de Desenvolupament Sostenible (Agenda 2030, ONU)

Diversitat i drets humans

Aquest treball està totalment alineat amb la dimensió de “Diversitat i drets humans”, concretament amb els objectius d’igualtat de gènere (ODS 5)⁵ i de “Reducció de les desigualtats (ODS 10)”⁶, amb la voluntat de “*Posar fi a totes les formes de discriminació en contra de totes les dones i nenes del món. (5.1)*” i “*Millorar l’ús de la tecnologia instrumental, en particular la tecnologia de la informació i les comunicacions, per promoure l’apoderament de la dona. (5.b)*”, més de “*Potenciar i promoure la inclusió social de les persones. (10.2)*” o “*Garantir la igualtat d’oportunitats i reduir les desigualtats de resultats. (10.3)*”.

Dimensió de comportament ètic i de responsabilitat social (RS)

Aquest treball té un impacte positiu en la dimensió de “Comportament ètic i de responsabilitat social”, ja que pot facilitar comportaments més equitatius, resolent situacions existents. Si més no, fent-se ressò de la problemàtica de l’ús d’una IA poc ètica. Esperant contribuir al bé comú de la societat, ajudant a reduir algunes desigualtats socials relacionades amb les discriminacions per raó de gènere.

Dimensió de sostenibilitat

Aquest treball es podria alinear amb l’objectiu “Indústria, innovació i infraestructures (ODS 9)”, procurant “*promoure una industrialització inclusiva i sostenible*” (9.2).

⁵ <https://www.un.org/sustainabledevelopment/gender-equality/>

⁶ <https://www.un.org/sustainabledevelopment/inequality/>

⁷ <https://www.un.org/sustainabledevelopment/es/infrastructure/>

En qualsevol cas, és un estudi que no té cap gran infraestructura al darrere, més enllà de l'ordinador personal i el consum d'energia de les 300 hores de treball. Tampoc es disposa dels recursos necessaris per mesurar la petjada ecològica d'aquest treball, tot i que per la creació dels models del llenguatge utilitzats en l'estudi s'utilitza una gran quantitat de recursos energètics. Així i tot, una vegada creats els models són fàcilment transportables i reutilitzables sense costos addicionals.

Perspectiva de gènere

Partint del punt que el treball en si **promou explícitament la igualtat de gènere** i la reducció de les desigualtats, s'ha procurat en tot moment no reforçar biaixos, estereotips ni preeminències masculines.

Sobre el llenguatge utilitzat en el treball, encara que no hi ha una normativa definida sobre el seu ús, s'ha procurat utilitzar un **llenguatge inclusiu**, fent servir indistintament tant l'ordre com les formes femenines com masculines. Quan s'anomenen professions amb marca de gènere gramatical, s'ha fet us de les dues formes juntes (bibliotecària/bibliotecari).

Finalment, malgrat que la realitat de gènere requereix un tractament multidimensional, en aquest estudi s'ha fet un **tractament binari del gènere**, **seguint** l'estructura i el gènere gramatical del català que també té una distinció binària. Tot i així, la realitat del gènere no és binaria, i aquest tractament binari podria excloure persones que s'identifiquen com no binàries i crear perjudicis de representació.

1.4. Enfocament i mètode seguit

Una de les majors problemàtiques que presentava el treball ha estat que degut al desconeixement previ en la matèria feia difícil una definició de l'abast. Amb aquesta dificultat, s'ha fet una definició prèvia de l'abast del projecte i una planificació inicial de les tasques a necessàries.

S'han dividit les tasques en **5 grups**:

1. Planificació del treball i detall de tasques a realitzar.
2. Estudi de l'estat de l'art del processament del llenguatge natural i dels models del llenguatge.
3. Estudi de l'estat de l'art sobre la detecció de biaixos en els models del llenguatge.
4. Adaptació de les metodologies de detecció i mesura de biaixos al model del llenguatge en català.
5. Anàlisi de resultats.

En l'estudi de l'estat de l'art del processament del llenguatge natural i dels models del llenguatge, s'ha comprovat que la **majoria de models del llenguatge són en anglès i una petita part es troba en altres llengües**. També s'ha estudiat

la tipologia, característiques i funcionament del model del llenguatge en català utilitzat en el treball.

De la mateixa manera, en l'estudi de l'estat de l'art sobre la detecció de biaixos en els models del llenguatge, s'ha evidenciat el domini de metodologies plantejades per l'anglès i les **dificultats d'adaptació d'aquests mètodes a altres idiomes**, com és el cas del català. Una de les principals diferències amb l'anglès és la **figura del gènere gramatical**, present en les llengües romàniques (com el català), d'articles, adjetius i noms que afecten la concordança de l'oració.

Tenint en compte aquestes dificultats, s'ha adaptat el mètode utilitzat per Bartl et al. (2020)[3], que analitza el grau d'associació entre un conjunt professions, amb el gènere representat per un conjunt de termes relatius a persones femenines (mare, germana, ella) i masculines (pare, germà, ella). L'adaptació d'aquest mètode al model en català, ha constat dels següents passos:

1. Recopilació de dades de professions de Wikidata amb els totals de dones i homes per professió.
2. Neteja i preparació de dades de professions.
3. Aplicació de la metodologia de Bartl a les professions seleccionades en el model del llenguatge en català.
4. Anàlisi de resultats i creació de taules i gràfics per la seva visualització i interpretació.

Programari emprat

Per aquest treball s'ha utilitzat el llenguatge de programació Python (v3.8.13)⁸, juntament amb les llibreries que es mostren a la **taula 2**.

Llibreria	Versió	Descripció
<i>SPARQLWrapper</i> ⁹	2.0.0	Recopilació de dades de Wikidata amb SPARQL ¹⁰ .
<i>Pandas</i> ¹¹	1.5.0	Anàlisis i manipulació de dades.
<i>Transformers</i> ¹²	4.20.1	Utilització de models preentrenats.

⁸ <https://www.python.org/>

⁹ <https://sparqlwrapper.readthedocs.io/en/stable/>

¹⁰ <https://www.w3.org/TR/sparql11-overview/>

¹¹ <https://pandas.pydata.org/>

¹² <https://huggingface.co/docs/transformers/>

<i>Torch</i> ¹³	1.12.1 Computació d'operacions matemàtiques amb estructures multidimensionals (Tensors) i suport de processament en GPU.
----------------------------	--------------------------------------------------------------------------------------------------------------------------

<i>Numpy</i> ¹⁴	1.23.4 Llibreria científica per efectuar operacions matemàtiques.
----------------------------	-------------------------------------------------------------------

Taula 2 - Llibreries de Python utilitzades.

El codi Python elaborat, de recopilació de dades, neteja i preparació de les dades, adaptació del mètode de Bartl i l'anàlisi de resultat s'ha penjat al repositori públic de GitHub ¹⁵.

Model utilitzat

El model del llenguatge en català analitzat en aquest treball, és el **model [4]** ¹⁶ **desenvolupat en el marc del projecte AINA** impulsat pel Departament de la Vicepresidència i de Polítiques Digitals i Territori del Govern de Catalunya, amb la col·laboració del Barcelona Supercomputing Center (BSC).

1.5. Planificació del treball

La planificació del treball ha estat una de les grans dificultats del treball. La manca de coneixement previ en la matèria i la complexitat d'alguns conceptes ha dificultat la planificació adequada de les tasques a desenvolupar així com de la definició de l'abast. Per aquest treball s'ha previst un total de 300 hores de treball, amb una desviació aproximada de 35 hores sobre el previst inicialment. A continuació es mostren les tasques dutes a terme en el treball, així com les hores previstes, les hores i la desviació¹⁷.

Tasca	Previst	Utilitzat	Desv.
Planificació	55	60	5
Definició i contextualització	20	25	5
Definició del projecte	10	10	
Contextualització del treball	10	15	5
Planificació	25	25	
Definició d'activitats i calendari	15	15	
Planificació de la gestió dels riscos	10	10	
Seguiment i control	10	10	
Reunions de seguiment i control de l'abast i qualitat	10	10	
Desenvolupament del projecte	180	195	-5

¹³ <https://pytorch.org/docs/stable/torch.html>

¹⁴ <https://numpy.org/>

¹⁵ <https://github.com/nurasaki/TFG-gender-bias>

¹⁶ <https://huggingface.co/projecte-aina/roberta-base-ca-v2>

¹⁷ Els valors de les hores utilitzades són orientatius.

Documentació i estudi de l'estat de l'art	50	45	-20
Estudi de tècniques de detecció i mesura de Biaixos	25	20	
Estudi de models del llenguatge de llengües no majoritàries (Català i altres)	20	10	-5
Estudi de detecció i mesura de Biaixos en llengües no majoritàries	15	10	-5
Estudi de tècniques de reducció de biaixos	10	0	-10
Revisió de les tasques planificades i afegir els canvis necessaris	5	5	
Aplicació de les tècniques de detecció de biaixos de gènere en els models seleccionats	130	150	5
Preparació de l'entorn de treball.	5	5	
Seleccionar models del llenguatge en Català.	5	5	
Estudiar les característiques tècniques i la tipologia dels models seleccionats.	15	20	5
Demanar assessorament a professors	10	5	-5
Generar dataset professions en català	15	25	10
Generar algorismes de creació de frases relacionant gènere i professions	20	20	
Aplicar/adaptar la metodologia amb els algorismes i dataset	25	15	-10
Estudi dels resultats obtinguts	20	30	10
Elaborar informe de resultats i conclusions	20	25	5
Documentació	65	80	25
Memòria	30	40	20
Agrupar tots els components del treball	15	20	5
Revisió i correcció de la memòria	15	20	5
Presentació i defensa	35	40	5
Agrupament i síntesi dels elements principals del treball	15	15	
Preparació de la presentació i la defensa	20	25	5
Total TFG	300	335	25

Taula 3 - Planificació de tasques

1.5.1 Anàlisi i gestió de riscos

A l'inici del projecte s'han definit els possibles riscos que poden afectar l'assoliment dels objectius del treball. Establir mesures preventives i correctives, a més de determinar la seva probabilitat i impacte tenir especial cura amb els riscos amb major rellevància. La **taula 4** mostra la taula els possibles riscos previstos inicialment amb el grau de probabilitat i el possible impacte.

Risc	Probabilitat	Impacte
Pèrdua d'informació o de material	Baix +	Alt
Manca de recursos de maquinari	Mig	Mig
Resultats poc clarificadors	Baix	Alt
Desviament de la planificació	Mig	Baix

Abast del projecte massa ampli	Mig	Alt
Falta de temps	Baix	Mig
Coneixements tècnics insuficients	Baix	Baix

Taula 4 - Anàlisi de riscos

La **pèrdua d'informació o de material** informàtic és un dels riscos més rellevants, i podria fer que algunes tasques assolides s'hagin de repetir. És important establir mesures preventives com una bona gestió de còpies de seguretat en discs externs i còpies al núvol. Es treballarà amb diversos repositoris de dades com Github (per emmagatzemar codi i la seva documentació) i Google Drive (documents, codi i models). Com a mesures correctives per la pèrdua de material, s'invertirà en la compra de nou material si fos necessari, amb la consegüent ampliació de costos del projecte.

Els models del llenguatge utilitzats en el treball, han sigut desenvolupats mitjançant xarxes neuronals, i per entrenar aquests s'han emprat una gran quantitat de **recursos de maquinari** (disponibles al Barcelona Supercomputing Center). Encara que fer per servir aquests models no es requereixen tants recursos, per tècniques com per exemple l'entrenament del model **podria ser necessària una ampliació de recursos**. Per tant, com a mesures preventives, no s'empraran tècniques ni es faran modificacions en els models que requereixin una gran quantitat de recursos de maquinari. Una mesura correctiva per aquesta limitació seria l'ampliació de recursos de maquinari amb l'increment de costos corresponents.

Un altre dels riscos, a tenir en compte és que **no s'aconsegueixin els resultats esperats**, o que no siguin representatius de la problemàtica que es vol tractar. En aquest sentit, caldrà investigar el perquè dels resultats i documentar-ho. Una de les possibles causes d'aquest risc, podria ser que les tècniques de detecció i mesura de biaixos actuals, majoritàriament estudiades per **models en anglès, no s'adapten als models del llenguatge català**. En aquest cas, com a mesura preventiva, caldrà realitzar un bon estudi de les tècniques existents de detecció de biaixos. Com a mesura correctiva es durà a terme una bona documentació dels motius.

Aquest projecte representa un repte desconegut, i inicialment no definició de tasques no és clara. Això pot comportar una probable **desviació de la planificació**. Com a mesura preventiva, serà necessari fer una bona segmentació de les activitats, i com a mesura correctiva s'implementaran metodologies àgils per adaptar la planificació als nous reptes que vagin sorgint. En aquest cas, serà important una bona comunicació amb el tutor per tal de redefinir fites i tasques. Pel mateix motiu, aquest repte desconegut pot fer que la definició de l'abast no s'ajusti al volum d'hores previstes per desenvolupar el projecte. Igual que el cas anterior, és necessari fer revisions continues de l'estat del projecte i redefinir l'abast en cas necessari. Per aquest treball està previst una inversió de 300 hores en 25 hores setmanals. Tractant-se d'un repte desconegut, serà necessari una bona gestió del temps per evitar el **risc de**

manca de temps. Com a mesura preventiva, serà necessari una bona planificació del projecte i definir un abast i objectius assumibles.

Finalment, les tecnologies utilitzades en aquest projecte, no s'ha estudiat en profunditat durant el grau, i, per tant, és un risc intentar assolir una fita de la qual encara **no es disposen de prou coneixement tècnic**. És important la fase de documentació i estudi dels models de llenguatge i tècniques de detecció de biaixos. Com a mesura preventiva serà necessari definir un abast i objectius assumibles. Com a mesura correctiva, serà necessari redefinir l'abast del projecte amb objectius assumibles segons els nostres coneixements.

1.6. Breu sumari de productes obtinguts

En aquest treball s'ha fet un estudi i anàlisi de l'existència de biaixos de gènere en un model del llenguatge basat en *deep learning* en català. S'ha adaptat un mètode inicialment definit per al llenguatge anglès i s'ha elaborat un informe detallat amb els resultats obtinguts pel model.

1.7. Breu descripció dels altres capítols de la memòria

A continuació fa una breu descripció de la resta de capítols de la memòria.

Al capítol 2, es fa un breu repàs del processament del llenguatge natural (PLN), en el marc de la intel·ligència artificial. S'analitzen les diferents representacions de paraules, així com els diferents models del llenguatge. També es fa un repàs de l'àtica en la IA i el PLN i s'analitzen els diferents mètodes de detecció de biaixos. Finalment, es detalla el funcionament del model emprat en el treball.

En el capítol 3 es detalla la metodologia seguida en el treball, amb la recollida, la neteja i la preparació de les dades que s'han utilitzat a l'estudi. També es detallen els passos seguits per l'adaptació del mètode de detecció de biaixos al model seleccionat.

Al capítol 4 es mostren els resultats aconseguits per les diferents professions seleccionades per l'estudi, això com la visualització i interpretació de resultats.

Finalment, al capítol 5 s'expliquen les conclusions obtingudes al treball, amb les lliçons apreses, els objectius assolits i les línies de futur obertes amb el present treball.

2. Processament del llenguatge natural

El processament del llenguatge natural PLN (*Natural Language Processing - NLP*) és una **branca de la intel·ligència artificial**, les ciències de computació i la lingüística que permet la **interacció entre ordinadors i persones a través del llenguatge humà**. El PLN té l'objectiu d'entrenar les màquines (ordinadors) per a què entenguin i generin llenguatge tal i com ho fan les persones.

En aquest capítol es fa un breu repàs del **PLN i les representacions de les paraules i els seus significats**, així com algunes de les tasques dutes a terme per sistemes de PLN. També es fa un repàs **dels reptes ètics i socials d'aquesta branca de la intel·ligència artificial**. Un d'aquests reptes és la detecció de biaixos, i és per això que també s'inclourà en aquest capítol la definició de biaixos i els tipus de biaixos que es poden trobar en el llenguatge, fent un repàs dels mètodes més importants per la detectar-los i mesurar-los.

2.1. Representació de paraules

La representació numèrica de les paraules i el seu significat, és el primer pas per tal que les màquines pugin “entendre” i processar les paraules. En els primers estadis del PLN, les primeres representacions numèriques de les paraules eren representacions discretes, com és la representació en vectors *one-hot* (**taula 4**). En aquesta representació, les paraules es representen amb vectors de dimensió igual a la mida del diccionari. Aquests vectors estan compostos per elements de 0 per a totes les posicions tret de la posició corresponent a la paraula que es representa amb el valor 1.

Paraula	Vector
Porta	[0, 0, 0, 1, 0, 0, 0, 0]
Finestra	[0, 0, 0, 0, 0, 0, 1, 0]

Taula 5 - Exemple de representació *one-hot* de dues paraules.

Aquest tipus de representació és d'utilitat quan la mida del vocabulari és petita i fixa, i és compatible amb models d'aprenentatge automàtic. Malgrat que aquesta representació és fàcil d'interpretar i d'implementar, aquest mètode té limitacions. Per una banda, quan la mida del vocabulari és molt alta, ens topem amb la “maledicció” de la dimensionalitat ¹⁸ (els vectors ¹⁹ són massa grans i genera costos molt elevats de computació). D'altra banda, i el que és més important, no es capture la informació semàntica i sintàctica de les paraules ni es representa les similituds entre elles.

2.2 Representació de paraules segons el context, els *Word embeddings*

“You shall know a word by the company it keeps” (Firth, J.R. 1957) [5].

Aquesta frase va ser una de les idees més revolucionàries del PLN, i fa referència a que el significat d'una paraula és definida per les **paraules que apareixen al seu costat**. Aquest principi proposa la representació de les paraules segons el

¹⁸ https://en.wikipedia.org/wiki/Curse_of_dimensionality

¹⁹ Un vector és un objecte matemàtic que s'utilitza en àlgebra vectorial, és necessari per fer operacions matemàtiques entre dos punts d'un espai multidimensional. Normalment, es representa amb una fletxa que uneix dos punts.

seu context (les paraules que l'acompanyen), i va ser la idea que va inspirar els **word embeddings** o vectors de paraules²⁰.

Els vectors de paraules són **representacions numèriques que capturen la informació semàntica i sintàctica de les paraules**, i va suposar una revolució en el món del processament del llenguatge natural. Els vectors de paraules s'utilitzen en multitud **d'aplicacions de PLN** i han suposat una millora en el rendiment de tasques com la traducció automàtica, l'extracció d'informació, la classificació de text o l'anàlisi de sentiments. Recentment, també han permès l'aparició i sofisticació dels xatbots o agents de conversació.

Existeixen diversos mètodes de creació de vectors de paraules:

- Mètodes basats en el [recompte de paraules](#)
- Mètodes basats en la [predicció de paraules](#)
- Mètodes basats en [aprenentatge profund o deep learning](#)

1. [Recompte de paraules](#): Els mètodes basats en el **recompte de paraules** analitzen la concorrència de paraules que apareixen al costat d'altres. Per exemple, si es defineix una finestra (*window context*) de 2 paraules, la paraula objectiu quedarà associada amb les 4 que té al voltant. En la següent frase:

“Les **plantes** del **jardí** són boniques.”

Jardí quedaria associat a “plantes” i a “boniques”, segons els algoritmes de recompte de paraules. GloVe [6] seria un exemple d'aquest mètode.

2. [Predicció de paraules](#): L'increment de la capacitat computacional, la generació de grans volums de dades de text i el desenvolupament de noves tecnologies com les xarxes neuronals va facilitar la creació de nous mètodes basats en la **predicció de paraules**, dissenyats per predir la següent paraula donat un context. *Word2vect* [7], és un algorisme basat en la predicció de paraules, publicat el 2013 per un equip de recerca de Google, que combina les arquitectures CBOW (*continuous bag of words*), i que intenta predir la paraula objectiu en funció de les paraules en un determinat context, i l'arquitectura *skip-gram* per predir les paraules que accompanyen (les paraules del context) una paraula d'entrada. Un altre exemple de mètode basats en la predicció de paraules és l'algorisme *FastText* desenvolupat per Facebook el 2014 [8].

Word2vect va ajudar a popularitzar la idea de la **representació de vectors de paraules**, i també la manera de com construir aquests vectors de paraules. Amb aquesta idea, es va comprovar que la representació en vectors de les paraules, va més enllà de les regles sintàctiques entre paraules. Aquestes representacions permeten efectuar operacions algebraiques o altres operacions matemàtiques per trobar similituds entre paraules. Per exemple, a la paraula "King" es resta el

²⁰ En aquest treball també ens referim a les representacions de *word embeddings* com a vectors de paraules.

vector de la paraula "man", que sumat al vector dona com a resultat la paraula "queen" [9].

La **figura 1** mostra la representació d'aquests vectors en un espai bidimensional, il·lustrat amb valors ficticis, mostrats a la **taula 4**. A la figura 1 es pot veure com les paraules "king" i "queen" són properes entre elles i, per tant, són paraules amb significats similars, de la mateixa manera que ho són les paraules "man" i "woman". El nombre de dimensions de les representacions reals amb els models obtinguts s'especifica com a paràmetre en l'algorisme word2vect, normalment amb vectors d'entre 300 i 1000 dimensions.

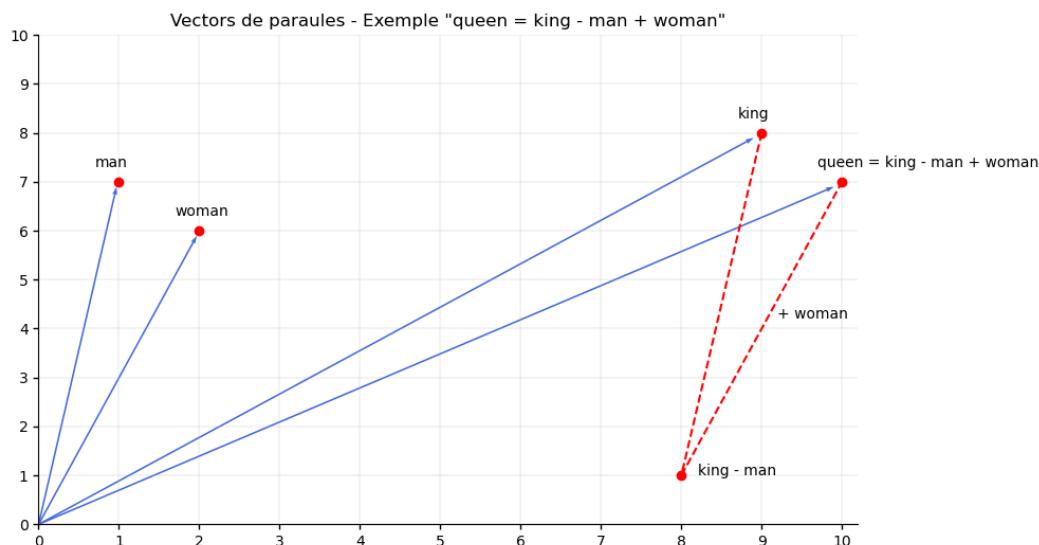


Figura 1 - Exemple de la representació de vectors de paraules 2D ($\text{queen} = \text{king} - \text{man} + \text{woman}$).

Paraula	Vector [x,y]
king	[9,8]
man	[1,7]
woman	[2,6]
queen	[10,7]
king - man	$[9,8] - [1,7] = [8,1]$
king – man + woman	$[9,8] - [1,7] + [2,6] = [10,7]$

Taula 6 - Exemple de representació bidimensional dels vectors "king-man-queen-woman".

Per mesurar la similitud entre vectors, es pot utilitzar la fórmula de la similitud del cosinus (**Figura 2**), amb el càlcul del producte escalar dels dos vectors A i B, dividit pel producte de les magnituds dels vectors o norma euclidiana. **El càlcul de similituds entre vectors de paraules s'utilitza en algun estudi com a mesura de biaix (programmer – man + woman = nurse).**

$$\text{cosine similarity} = S_C(\mathbf{A}, \mathbf{B}) := \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}},$$

Figura 2 - Similitud del cosinus ²¹

3. *Aprendentatge profund o deep learning*: En la secció 2.3 es fa una explicació més detallada dels mètodes basats en **deep learning**.

2.3 Word embeddings i deep learning

La millora del maquinari i l'optimització dels algorismes de vectors de paraules basats en deep learning representen l'estat de l'art actual dels models del llenguatge. Una bona representació de paraules ha de tenir en compte les **complexitats i la naturalesa semàntica i grammatical de les paraules**. A més, de com varien en funció del context que acompanya les paraules.

Una de les primeres representacions que va tenir en compte el context variable de les paraules, a diferència de les representacions estàtiques dels *word embeddings* tradicionals, és el model ELMo [Peters], que processa seqüències de dades com a valors d'entrada utilitzant una arquitectura LSTM (*Long short-term memory*) bidireccional. Aquest és un mecanisme de xarxa neuronal recurrent (RNN) que té en compte les seqüències de dades anteriors i posteriors.

Posteriorment, OpenAI²² va desenvolupar GPT [10] precursor de GPT-2, GPT-3, i de ChatGPT, que utilitza l'arquitectura del *Transformers* [11], també conegut com a mecanismes d'atenció per extreure les característiques de les paraules. A diferència de les anteriors RNN, que en cada nou estat de la xarxa només té en compte el darrer estat del codificador, els mecanismes d'atenció tenen en compte tots els estats anteriors. Així, permeten assignar major pes o importància a certs elements de la seqüència d'entrada, que es podria entendre com que la xarxa presta més atenció a certes parts del text que a d'altres.

Un altre dels models basats en *deep learning* de referència és el **model BERT** (*Bidirectional Encoder Representations from Transformers*)[12], que té una arquitectura de xarxa amb un Transformer bidireccional i utilitza un mètode d'aprenentatge autosupervisat com a mètode d'entrenament. Aquest mètode s'anomena **modelat del llenguatge emmascarat** (*Masked Language Modeling* - MLM), que consisteix a emmascarar (substituir per "[MASK]") aleatoriament el 15% de les paraules i posteriorment predir quines són les paraules emmascarades. Aquesta tasca de MLM s'empra en aquest treball per la predicció de paraules susceptibles de contenir biaixos.

²¹ https://en.wikipedia.org/wiki/Cosine_similarity

²² <https://openai.com/>

Exemple d'emmascarat aleatori del 15% de les paraules:

A father and his son are in a [MASK] accident. The father dies at the [MASK] and the son is rushed to the [MASK]. At the hospital the [MASK] [MASK] at the boy and says "I can't [MASK] on this [MASK], he is my [MASK]."

How can this be?

El desenvolupament de BERT, i les posteriors variants com RoBERTa (Robustly Optimized BERT Pretraining Approach) [13] va suposar un gran canvi obrint les portes al **transfer learning**, que permet reutilitzar models preentrenats per altres tasques i idiomes (**taula 5**). Aquesta funcionalitat, juntament amb tècniques autosupervisades, va fer possible que els llenguatges amb pocs recursos (poques dades de text per entrenar els models) poguessin beneficiar-se dels grans avenços assolits en anglès.

Tasca NLP

Classificació de textos (Text Classification - TC)

Pregunta resposta (Question Answering - QA)

Similitud semàntica (Semantic Textual Similarity - STS)

Etiquetats de paraules (Part-Of-Speech tagging - POS)

Reconeixement d'entitats nominals (Named Entity Recognition - NER)

Classificació de textos (Text Classification - TC)

Taula 7 - Tasques NLP utilitzades en transfer learning.

2.3.1 BERTa, el model català

El *transfer learning* i la reutilització del model BERT, ha permès la creació de nous models multillenguatge, com mBERT²³. Aquest model entrenat per 104 llengües, inclòs el Català, aprofita les dades de la Viquipèdia. De la mateixa manera, el model RoBERTa té la seva versió multilingüe amb XLM-RoBERTa [14] entrenat per 100 llenguatges a partir de 2,51Tb de dades obtingudes de CommonCrawl²⁴.

Malgrat que els models mBERT i XLM-RoBERTa obtenen bons resultats, els models monolingües entrenats amb dades pròpies obtenen millor rendiment. És el cas del model francès CamemBERT [15], l'holandès BERTje [16], o el finès FinBERT [17]

Les causes de la millora d'aquests models específics, són que utilitzen una quantitat i diversitat de dades més gran. També, pel fet de tenir un vocabulari

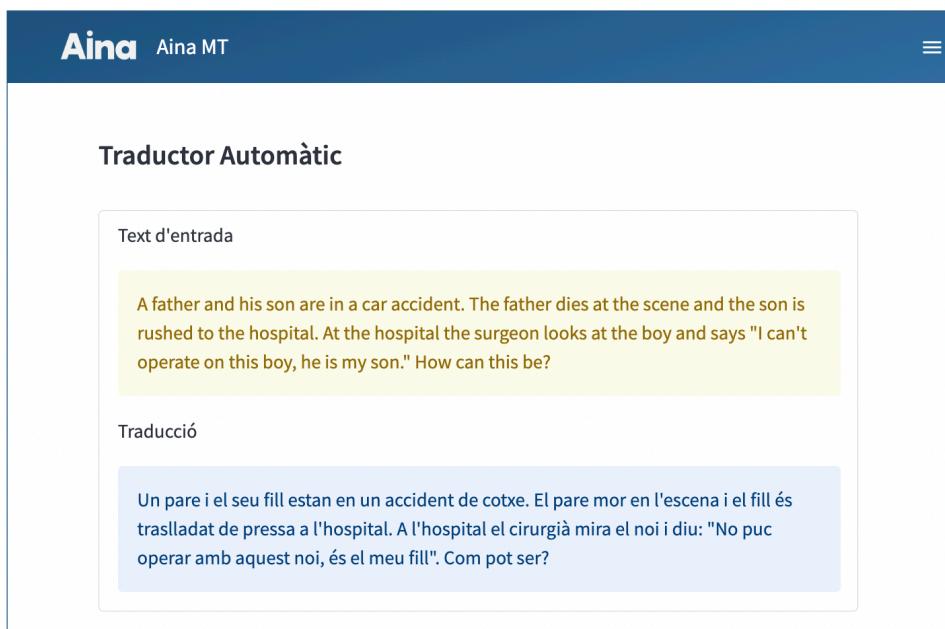
²³ <https://huggingface.co/bert-base-multilingual-cased>

²⁴ <https://commoncrawl.org/>

específic prop del llenguatge, evita la divisió de les paraules que no es troben al vocabulari del model en subparaules, fent-lo lingüísticament menys interpretable. Aquests models tenen un límit en la mida del vocabulari. En el cas del model BERT (i derivats) és d'unes 30.000 paraules i subparaules, i en el cas de RoBERTA és d'unes 50.000 paraules. Aquest aspecte s'explicarà amb més detall a l'apartat 2.3.4 "Tokenització".

El projecte AINA

El 2020 el Departament de la Vicepresidència i de Polítiques Digitals del Govern de Catalunya, va presentar el projecte AINA²⁵ amb l'objectiu de “proveir el català de la infraestructura per al desenvolupament d'aplicacions basades en AI”. Amb l'ajuda del Barcelona Supercomputing Center (BSC) s'han generat diferents recursos, com grans volums de dades textuales netejades i models del llenguatge. També s'han desenvolupat algunes eines, com un xatbot català o traductors automàtics²⁶.



Il·lustració 4 - Traducció endevinalla cirurgià - Aina MT. Consulta: 22-12-2022. ²⁷

En aquest context, el 2021, aprofitant el model RoBERT, es **va crear el model BERTa [4]** amb dades de text en català de diferents fonts, com la Viquipèdia, l'agència Catalana de Notícies o documents del Butlletí Oficial de la Generalitat de Catalunya. La **il·lustració 5**, extreta de l'article original de BERTa [4], mostra les mètriques de rendiment superiors a les obtingudes en els models multillenguatge mBERT i XLM-RoBERT, per les diferents tasques (**Taula 5**). Aquestes mètriques de rendiment del model avaluen la seva resposta comparantlo únicament a nivell de dades, és a dir, com més semblants siguin les respostes del model a les dades, millors resultats. L'utilització d'aquestes

²⁵ <https://politiquesdigitals.gencat.cat/ca/economia/catalonia-ai/aina/>

²⁶ <https://aina.bsc.es/>

²⁷ <https://traductor.aina.bsc.es/>

mètriques tradicionals comporta un problema si les dades utilitzades contenen algun tipus de biaix, ja que els models “aprenen”, propaguen i, fins i tot, amplifiquen [1] aquest biaixos.

Les autors i autors d'aquests models són conscients d'aquesta problemàtica i fan una declaració a l'article [4]: “***we hypothesize that the pre-training corpus will have different biases and the model might reproduce them, so users must be aware of this issue.***”

model	NERC	POS	STS	TC	QA (ViquiQuAD)	QA (XQuAD)
BERTa	88.13 (2)	98.97 (10)	79.73 (5)	74.16 (9)	86.97/72.29 (9)	68.89/48.87 (9)
+ decontaminate	89.10 (6)	98.94 (6)	81.13 (8)	73.84 (10)	86.50/70.82 (6)	68.61/47.26 (6)
mBERT	86.38 (9)	98.82 (9)	76.34 (9)	70.56 (10)	86.97/72.22 (8)	67.15/46.51 (8)
WikiBERT-ca	77.66 (9)	97.60 (6)	77.18 (10)	73.22 (10)	85.45/70.75 (10)	65.21/36.60 (10)
XLM-RoBERTa	87.66 (8)	98.89 (10)	75.40 (10)	71.68 (10)	85.50/70.47 (5)	67.10/46.42 (5)

Table 4: Results for the downstream tasks using different metrics. We use F1 for POS and NERC, accuracy for TC, an average of Pearson and Spearman coefficient for STS and F1/Exact Match for QA. We also report within round brackets the best epoch on the dev set.

II·I·lustració 5 - Mètriques de rendiment model BERTa [4]

Disclaimer

▼ Click to expand

The models published in this repository are intended for a generalist purpose and are available to third parties. These models may have bias and/or any other undesirable distortions.

When third parties, deploy or provide systems and/or services to other parties using any of these models (or using systems based on these models) or become users of the models, they should note that it is their responsibility to mitigate the risks arising from their use and, in any event, to comply with applicable regulations, including regulations regarding the use of Artificial Intelligence.

In no event shall the owner and creator of the models (BSC – Barcelona Supercomputing Center) be liable for any results arising from the use made by third parties of these models.

II·I·lustració 6 - Exempció de responsabilitat a la pàgina d'Hugging Face de BERTa.²⁸

²⁸ <https://huggingface.co/projecte-aina/roberta-base-ca-v2>

2.4. Ètica, intel·ligència artificial i PLN

L'aplicació d'algorismes **d'aprenentatge automàtic (Machine Learning)** i la intel·ligència artificial es troben en multitud d'aplicacions de presa de decisions que utilitzem en el nostre dia a dia. A conseqüència d'això, cada cop més, es presenten escenaris d'alt risc com ara la presa de decisions automàtiques sobre préstecs bancaris [18] o la contractació de personal [19]. Com s'ha comentat anteriorment, sovint, aquestes aplicacions **no estan lliures de biaixos i la seva utilització podria comportar danys socials** que s'ha de treballar per prevenir.

En un dels casos que ha tingut més ressò, l'any 2016 periodistes de ProPublica van mostrar l'existència de pràctiques discriminatòries i racistes amb l'aplicació de l'eina COMPAS²⁹ (*Correctional Offender Management Profiling for Alternative Sanctions*), que utilitza un model d'aprenentatge automàtic per predir la probabilitat de reincidència de determinats presos i preses [20].

Aquesta problemàtica ja ha sigut tractada per òrgans legislatius. A la UE és aplicable des de 2018 la regulació general de protecció de dades GDPR³⁰, que introduceix conceptes com equitat i transparència. O la Llei d'intel·ligència artificial, que **regula l'aplicació d'eines basades en intel·ligència artificial d'alt risc** (sistemes de predicció de crims, o per reforçar forces judicials i policials), que tenen l'obligació d'aportar informació sobre la validació de les dades, les mètriques emprades i els impactes potencialment discriminatoris.

Definició de biaix

Segons el Termcat, un biaix de gènere és el "**Resultat d'un plantejament erroni d'igualtat o de diferències entre dones i homes que causa una situació desigual i discriminatòria per a un gènere respecte de l'altre. És un exemple de biaix de gènere la Infrarepresentació d'un gènere i la suprarepresentació de l'altre en un àmbit determinat.**"³¹.

En el context dels models del llenguatge emmascarats com són BERT o RoBERTa, un biaix és la tendència del model a retornar valors de sortida favorables o desfavorables per a un grup determinat, com ara homes o dones, donat el mateix context emmascarat [21].

Dins dels sistemes d'IA podem trobar **4 tipus** de biaixos:

1. Denigració (tractament ofensiu envers un grup)
2. Estereotipat (percepció errònia que un grup comparteix unes certes característiques).
3. Biaix de reconeixement (quan el sistema no reconeix un grup)
4. Infrarepresentació (quan un grup no es representa adequadament en el sistema)

²⁹ <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

³⁰ <https://gdpr-info.eu/>

³¹ <https://www.termcat.cat/es/diccionaris-en-linia/285/fitxa/MzY5NzU1MQ%3D%3D>

La **il·lustració 7** mostra com la traducció de Google Translate fa traduccions estereotipades de les professions introduïdes. Per exemple, tradueix “surgeon” per cirurgià en comptes de per cirurgiana.

The screenshot shows a comparison between English and Catalan using Google Translate. The English side lists terms like "Non-inclusive language", "The school children", "Stereotypes", "The surgeon", "The nurse", "The doctor", "The school teacher", "The kindergarten teacher", and "The university teacher". The corresponding Catalan translations include "Llenguatge no inclusiu", "Els nens de l'escola", "Estereotips", "El cirurgià", "La infermera", "El metge", "La mestra de l'escola", "La mestra d'infantil", and "El professor universitari". A red star icon is visible next to the last translation.

English	Catalan
=> Non-inclusive language	=> Llenguatge no inclusiu
The school children	Els nens de l'escola
=> Stereotypes	=> Estereotips
The surgeon	El cirurgià
The nurse	La infermera
The doctor	El metge
The school teacher	La mestra de l'escola
The kindergarten teacher	La mestra d'infantil
The university teacher	El professor universitari

Il·lustració 7 - Biaixos en Google Translate. Consulta 10-01-2023

2.4.1. Mesura de biaixos de gènere en els models del llenguatge

En els darrers anys, juntament amb la proliferació dels models del llenguatge i les tecnologies de PLN, hi ha hagut un increment en els estudis de biaixos de gènere en els models del llenguatge. La **taula 6** mostra els Articles publicats a ACL Anthology³² amb els termes “gender bias” al títol.

Any	Publicacions ACL
2016	1
2017	1
2018	3
2019	20
2020	23
2021	30

Taula 8 - Articles publicats a ACL Anthology amb “gender bias” al títol.

³² <https://aclanthology.org/>

Els primers mètodes de detecció i mesura de biaixos mostren com els *word embeddings* tradicionals capturen biaixos i estereotips presents a les dades de text [22], [23], [24]. Aquests mètodes basats en analogies de paraules, mostren com aquestes analogies capturen aquests biaixos estereotipats amb analogies com '***man is to computer programmer as woman is to homemaker***' [22]. Tot i que aquestes analogies són fàcilment interpretables, aquest mètode pot presentar algunes limitacions, ja que està basat en eleccions subjectives de paraules [25].

Posteriorment, Caliskan et al. (2017) [24] crea el test d'associació per als *word embeddings* (*Word Embedding Association Test - WEAT*). Aquest mètode, amb el càlcul de la similitud del cosinus entre els vectors (**Figura 2**), mesura l'associació entre paraules relatives al gènere (per exemple, noms de dona i de home) i paraules atribut (bo/dolent). Amb aquest mètode es va comprovar que els vectors de paraules de *GloVe* [6] contenen biaixos de gènere i de raça.

De la mateixa manera, també es va comprovar que aquests biaixos es propagaven en altres tasques de PLN. Per exemple, en tasques d'anàlisi de sentiments, es va observar que algunes mesures obtenien resultats diferents en funció de la raça o del gènere [26].

Word Embeddings basats en deep learning

Els *word embeddings* basats en deep learning estan **basats en els mètodes dels word embedding tradicionals**, que són els primers mètodes de detecció de biaixos que va haver. L'adaptació dels mètodes utilitzats en els *word embeddings* tradicionals, mostren com els **vectors de paraules contextualitzats també codifiquen els biaixos**, encara que no són massa evidents [27].

En els *word embeddings* basats en deep learning (també anomenats *word embeddings* contextualitzats), les representacions de vectors de paraules no són estàtiques sinó que **depenen de les frases que accompanyen cadascuna de les paraules (context)**, i per tant no es poden aplicar adequadament la majoria de mètodes anteriors basats en les similituds i projeccions dels vectors. Per això, alguns/es autors/es [28] [29] utilitzen **mètodes basats en plantilles per generar frases o contextos** que capturin les dinàmiques dels vectors.

Kurita et al (2019) [29], aprofita el modelatge del llenguatge emmascarat o *Masked Language Modeling* (MLM) utilitzat per entrenar el model BERT [12]. Amb aquest mètode obté la **probabilitat d'una paraula objectiu emmascarada que denoti el gènere** d'anar accompanyada d'un **context amb un cert atribut estereotipat** (per exemple una professió). Kurita determina l'associació entre la paraula "objectiu" (gènere) i l'atribut comparant les probabilitats de la paraula del gènere amb l'atribut emmascarat i l'atribut sense emmasclarar. L'estudi mostra com el **model BERT té major tendència a associar els atributs mesurats amb el gènere masculí** (aproximadament el 80% de les proves), evidenciant el biaix de gènere que presenta el model BERT.

Posteriorment, Bartl et al. (2020) [3] utilitza el mètode de Kurita ampliant el nombre de paraules de gènere i les plantilles per generar més varietat de frases. Per **mesurar l'associació entre el gènere i un conjunt de professions amb estereotip de gènere** extretes de l'oficina d'estadística del treball dels EUA³³. Els resultats mostren com les **professions amb estereotip masculí, estan fortemet associades a les paraules amb gènere masculí**, mentre que les professions amb **estereotip femení, estan fortemet associades a les paraules amb gènere femení**.

En el mateix estudi de Bartl, també s'estudia el mètode en el model alemany adaptat de BERT, però donat que l'alemany conté marques de gènere no s'obtenen resultats, ja que en aquest cas el **gènere gramatical té una forta influència en les associacions entre la paraula objectiu (gènere) i les professions**. Aquest mètode és que s'utilitzarà en el present treball i està detallat en el Capítol 3.

No s'han trobat altres estudis similars per a models en llengües amb marcat de gènere per adaptar en aquest estudi.

2.5. Funcionament dels models tipus BERT

2.5.1. Tokenització.

Els models del llenguatge utilitzats en **tasques de processament del llenguatge natural**, no poden entendre dades textuals en cru (paraules), sinó que s'han de convertir en dades numèriques (nombres) perquè puguin ser processades. Aquest procés s'anomena "tokenització"³⁴, i consisteix en convertir les dades de text en dades numèriques.

En el **procés de tokenització**, primer es divideix el fragment de text en unitats més petites anomenades *tokens*. Posteriorment, a cadascun d'aquests *tokens* se li assigna un identificador numèric, convertint així la cadena de text en una seqüència numèrica que s'utilitzarà com entrada al model.

Hi ha diferents mètodes de *tokenització*, que depenen de les necessitats del model. Els tres tipus d'algorismes més destacats, són:

- Algorismes basats en la divisió per **paraules**
- Algorismes basats en la divisió per **caràcters**
- Algorismes basats en la divisió per **subparaules**

³³ <https://www.bls.gov/cps/cpsaat11.htm>

³⁴ S'ha traduït de la paraula manlevada de l'anglès *tokenization* a falta d'un terme en català.

1. Divisió per paraules

Els algorismes basats en la divisió per paraules consisteixen, com el mateix nom indica, en la divisió per paraules, separades generalment per espais o per altres signes de puntuació. Per exemple, la frase “La Maria és infermera!”, es podria dividir en “La”, “Maria”, “és”, “infermera”, “!”. A cadascuna de les paraules li correspondría un identificador que si se substitueix per la paraula corresponent, podria quedar d'aquesta manera: [34, 675, 3343, 26, 3]³⁵.

Tot i que la informació emmagatzemada a cada unitat numèrica és alta (conté informació semàntica i del context), aquest algorisme té alguns inconvenients. Per exemple, per a poder modelar l'elevat nombre de paraules que conté una llengua, s'ha de crear un vocabulari molt extens amb l'identificador de cadascuna, fent que els recursos de computació necessaris siguin molt elevats i poc eficients. Una possible solució és limitar-ho a un nombre determinat de paraules, creant un vocabulari amb les N paraules més freqüents. En aquest cas, les paraules que no es trobin al vocabulari (*out-of-vocabulary words*) seran catalogades com “desconeegudes”, amb la pèrdua d'informació corresponent.

2. Divisió per caràcters

En els algorismes de divisió per caràcters, es divideix la cadena de text en una seqüència de caràcters i posteriorment en el seu identificador numèric. Així, la frase “La Maria és infermera.”, es podria dividir en 'L', 'a', 'M', 'a', 'r', 'i', 'a', 'é', 's', 'i', 'n', 'f', 'e', 'r', 'm', 'e', 'r', 'a', '!', i la possible correspondència numèrica [76, 97, 77, 97, 114, 105, 97, 233, 115, 105, 110, 102, 101, 114, 109, 101, 114, 97, 46]. Aquest algorisme té l'avantatge que amb només 256 *tokens* es poden representar la majoria de caràcters alfanumèrics.

Aquest mètode, però, també té alguns inconvenients. En primer lloc, la informació emmagatzemada en cada *token* és menor que la que pot contenir una sola paraula. El caràcter “a” no emmagatzema tota la informació que pot contenir la paraula “infermera”. Per altra banda, com es pot veure en la frase d'exemple, les frases d'entrada es transformen en cadenes molt llargues per ser processades pel model i, a causa de la limitació en el nombre d'elements d'entrada, es redueix la mida de les cadenes que poden ser utilitzades.

3. Divisió per sub-paraules

La *tokenització* basada en subparaules (*subword-based tokenizers*), que està entremig dels dos mètodes explícits anteriorment, és l'algorisme utilitzat pels models BERT, RoBERTa o BERTa, emprats en el present estudi. En aquest mètode, les paraules més freqüents es mantenen com a paraules senceres i no es divideixen en subparaules, mentre que les paraules menys freqüents es descomponen en subparaules.

En el cas del model en català BERTa, la paraula “infermera” és una paraula freqüent que no es divideix en subparaules, i que el model *tokenitza* amb l'identificador 26849. En canvi, la paraula “infermer”, molt menys freqüent, es

³⁵ Els identificadors de cada paraula són inventats, per il·lustrar l'exemple.

divideix en les subparaules “infer” (id 4139) i “mer” (id 4653) dotades de cert significat. Aquest exemple representa un cert biaix, ja que la resposta del model és diferent en funció del gènere, i en el present estudi no es podrà analitzar el comportament del model per la professió “infermer/infermera”, donat que s’han seleccionat professions representades amb un sol token.

Existeixen diversos algorismes de tokenització basats en subparaules. El model BERT utilitza l’algorisme WordPiece [30] desenvolupat per Google, i els models RoBERTa i BERTa utilitzen l’algorisme BPE (Byte-pair Encoding) [31]. La descripció detallada del funcionament d’aquests algorismes queda fora de l’abast d’aquest treball, però ambdós algorismes són molt similars i consisteixen en una divisió inicial per caràcters i posteriorment va agrupant per les parelles de combinacions més freqüents ampliant a combinacions superiors obtenint les paraules i subparaules més freqüents. Durant la creació dels models es defineix la mida del vocabulari, en el cas de BERT conté 30.522 tokens, RoBERTa 50.265 i el model català BERTa en té 50.262.

2.5.2. Funcionament dels models BERT i RoBERTa i la tasca fill-mask.

A continuació es fa una breu explicació del funcionament dels valors obtinguts pels models BERT [12], RoBERTa [13] i la seva versió en català BERTa [4] en la tasca “fill-mask” (omplir tokens emmascarats) ³⁶.

Seguint amb el model BERTa, les següents paraules/símbols es corresponen amb un token específic, és a dir una posició en diccionari, que serà sempre el mateix.

Paraula	Token ID
Ell	7785
Ella	11947
és	423
actor	10740
actriu	13751
advocat	11163
advocada	32308
infermera	26849
<mask>	4
	2462

Taula 9 - Exemple relació paraules - tokens (BERTa)

³⁶ [https://github.com/nurasaki/TFG-gender-bias/blob/main/Exemples/Exemple%20-%20Funcionament%20BERTa%20\(fill-mask\).ipynb](https://github.com/nurasaki/TFG-gender-bias/blob/main/Exemples/Exemple%20-%20Funcionament%20BERTa%20(fill-mask).ipynb)

Quan es vol analitzar la resposta del model, introduint una frase emmascarada, abans s'ha de *tokenitzar* (convertir les paraules en el seu equivalent numèric) perquè pugui ser entesa pel model. També s'afegeixen els *tokens* “<s>” (id 0) i “</s>” (id 2), que són caràcters especials per especificar l'inici i final de les frases.

Durant l'execució de la tasca “*fill de mask*”, s'introdueix una frase emmascarada (prèviament tokenitzada perquè el model la pugui llegir). Per exemple, la frase “Ella és <mask>.” *Tokenitzada* és equivalent a la seqüència numèrica [0, 11947, 423, 4, 2462, 2].

Posició	0	1	2	3	4	5
Paraula	<s>	Ella	és	<mask>	.	</s>
Token ID	0	11947	423	4	2462	2

Taula 10 - Tokenització per al model BERTa de la frase “Ella és <mask>.”

Per tant, si introduïm al model el vector numèric equivalent a la frase “Ella és <mask>.”, el model retornarà una matriu de valors de mida <mida vector entrada> x <mida vocabulari>. En aquest cas el vector d'entrada té una mida de 6, i el vocabulari té una mida de 50262, per tant, la mida de la matriu de sortida és de 6 x 50262. Els valors retornats es normalitzen amb la funció *softmax* per obtenir les probabilitats.

Posició (token ID)	Paraula	Valor (logits)	Valor (probabilitat)
0	<s>	-3.73	0.00000
1	<pad>	-6.42	0.00000
...
13751	actriu	4.38	0.00241
13752	abraçada	-0.62	0.00002
...
26848	constat	-4.76	0.00000
26849	infermera	5.09	0.00491
26850	perseguit	-4.18	0.00000
...
32308	advocada	3.50	0.00100
32309	cursat	-5.03	0.00000
...
50261	rigadista ³⁷		-6.43
			0.00000

Taula 11 - Valors retornats per el model BERTa per la posició 3 i la frase “Ella és <mask>.”

³⁷ El token id 50261 es correspon amb la subparaula “rigadista” (no és una paraula). Per exemple, la paraula “brigadista” es tokenitzaria amb [402, 50261]. L'id 402 es correspon amb el token “b” i l'id 50261 amb el token “rigadista”.

En aquest cas, si observem les posicions de les paraules “infermera”, “actriu” i “advocada”, obtenim unes probabilitats de 0.0049, 0.0024 i 0.0010 respectivament. La paraula “infermera” és el doble de probable que “actriu”, que alhora és més del doble que la paraula “advocada”.

La **figura 3** mostra l’esquema del funcionament de BERTa per la tasca *fill-mask*.

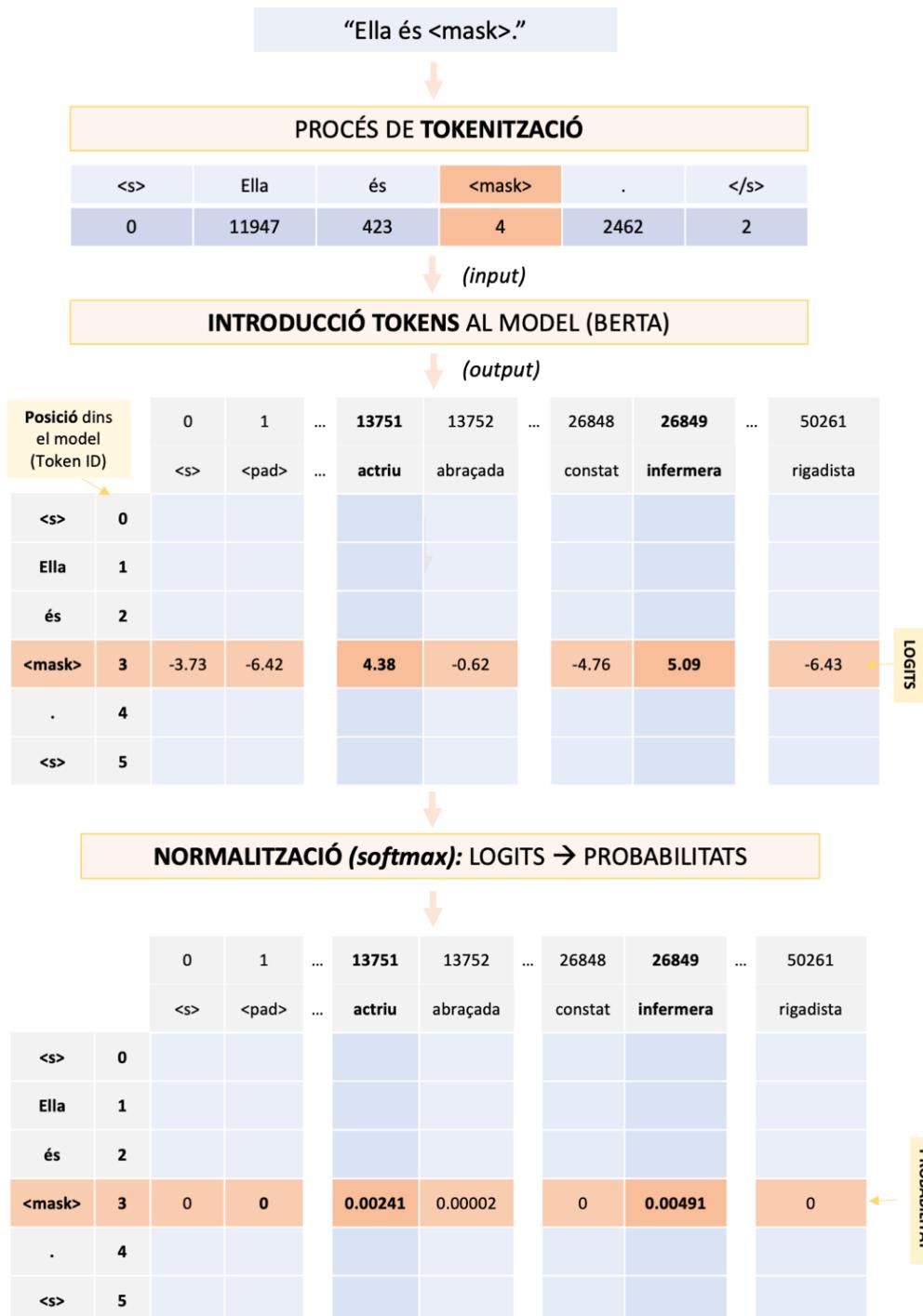


Figura 3 - Funcionament BERTa, tasca *fill-mask*.

3. Materials i mètodes

En aquest capítol es detallarà la metodologia utilitzada per la **detecció de biaixos en el model català BERTa**.

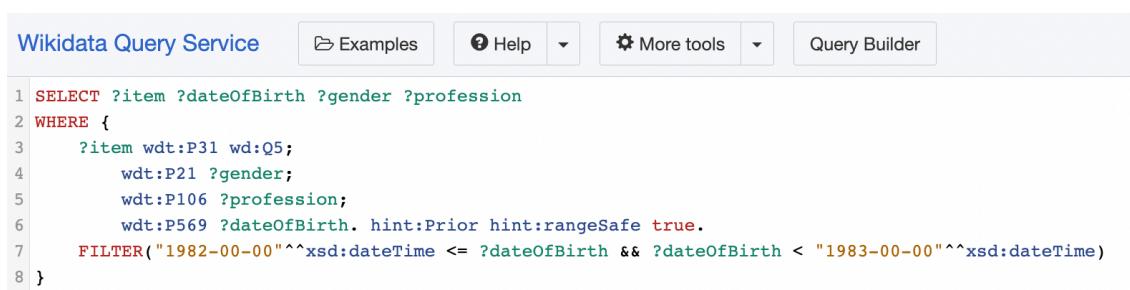
Els principals passos que s'han seguit són els següents:

1. Recol·lecció de dades de professions
2. Neteja de dades de professions
3. Adaptació del mètode de Bartl.
4. Obtenció dels valors d'associació gènere – professió.

3.1 Recol·lecció de dades de professions

Per tal de reproduir el mètode de Bartl et al. (2020) [3], que utilitza dades de l'Oficina del Treball dels EUA, es van demanar les dades de percentatges de treballadors i treballadores (home/dona) per professions tant a l'Institut d'Estadística de Catalunya (IDESCAT) com a l'Institut Nacional d'Estadística (INE)³⁸, però no es van proporcionar.

Com a alternativa, s'han recopilat les dades de professions a partir de les dades de *Wikidata*³⁹, que és una base de dades lliure i fa d'emmagatzematge central de dades estructurades per projectes com ara Viquipèdia, i l'eina *Wikidata Query Service*⁴⁰. Aquesta eina permet obtenir les dades amb llenguatge estàndard de consulta i manipulació de bases de dades SPARQL⁴¹. D'aquesta manera, s'han pogut recopilar les professions i gènere de totes les persones (amb entrada a la Viquipèdia) nascudes a partir de l'any 1900. Per exemple, amb la consulta mostrada a la **Il·lustració 8** s'obtenen les instàncies (P31) de personnes (Q5), el gènere (P21), la professió (P106) i la data de naixement (P569), nascudes l'any 1984⁴².



The screenshot shows the Wikidata Query Service interface. At the top, there are tabs for 'Examples', 'Help', 'More tools', and 'Query Builder'. The main area contains a SPARQL query:

```
1 SELECT ?item ?dateOfBirth ?gender ?profession
2 WHERE {
3   ?item wdt:P31 wd:Q5;
4     wdt:P21 ?gender;
5     wdt:P106 ?profession;
6     wdt:P569 ?dateOfBirth. hint:Prior hint:rangeSafe true.
7 FILTER("1982-00-00"^^xsd:dateTime <= ?dateOfBirth && ?dateOfBirth < "1983-00-00"^^xsd:dateTime)
8 }
```

Il·lustració 8 - Consulta SPARQL a *Wikidata Query Service*.

³⁸ <https://www.ine.es/>

³⁹ <https://www.wikidata.org/>

⁴⁰ <https://query.wikidata.org/>

⁴¹ <https://www.w3.org/TR/rdf-sparql-query/>

⁴² https://bit.ly/wikidata_persones_1984

Així, el procés que s'ha seguit per la recollida de professions i gènere és el següent:

1. Obtenció de persones, gènere, professió i data de naixement
2. Filtre per gènere masculí (Q6581097) i femení (Q6581072)
3. Agrupament de dades segons la professió
4. Obtenció dels noms de les professions en català i en anglès.

Per aquesta tasca, s'ha fet servir principalment la llibreria Pandas⁴³ (v1.5.0), que és una eina que facilita l'anàlisi i manipulació de dades, i la llibreria SPARQLWrapper (v2.0.0), que permet fer consultes SPARQL amb Python.

En total, **s'han recopilat 10.553 professions**, etiquetades en català o en anglès, que inclouen el total d'homes i dones.

3.2 Neteja i preparació de dades de professions

Després de la recopilació de les dades de *Wikidata*, s'han dut a terme diverses **tasques de pre-processat i neteja** de dades:

1. Eliminació de professions amb ≤ 10 personnes.

Aquelles professions amb menys de 10 persones, com per exemple “cercador de tresors” (Q17062828, amb 4 persones⁴⁴), han estat eliminades donada la seva baixa representació a la població.

2. Eliminació de professions sense etiqueta.

Algunes de les 10.553 professions no estan etiquetades en català (com per exemple “japanologist”, Q15991218⁴⁵) i s'han eliminat.

3. Revisió i modificació manual de les dades obtingudes.

A les dades extretes de *Wikidata*, s'han fet algunes modificacions manualment, com afegir els termes femenins de les professions. Alguns termes coincideixen amb el terme masculí, com per exemple taxista. Altres formes femenines que accepten dos termes, com per metge/metgessa o jutge/jutgessa, s'ha optat pel terme únicament femení, metgessa i jutgessa en aquests casos.

4. Agrupació de professions del mateix grup.

Algunes professions d'un mateix grup, per exemple “escriptor” i “escriptor de literatura infantil”, s'han unificat en una sola. A la **taula 12** es pot veure les professions agrupades en escriptor/escriptora. Els totals d'homes i dones resultant, és la suma de totes les professions.

⁴³ <https://pandas.pydata.org/>

⁴⁴ <https://www.wikidata.org/wiki/Q17062828>

⁴⁵ <https://www.wikidata.org/wiki/Q15991218>

	Etiqueta català	Wikidata ID	Dones	Homes	Total
1	escriptor	Q36180	58.064	126.111	184.175
2	escriptor de literatura infantil	Q4853732	4.182	3.527	7.709
3	escriptor de no-ficció	Q15980158	1.360	4.960	6.320
4	assagista	Q11774202	1.360	3.766	5.126
5	escriptor de ciència-ficció	Q18844224	972	3.529	4.501
6	prosista	Q12144794	685	2.471	3.156
7	escriptor de contes	Q15949613	548	728	1.276
...
24	escriptor de jardineria	Q100983844	7	6	13
escriptor / escriptora			68.101	147.372	215.473

Taula 12 - Professions Wikidata agrupades en actor/actriu.

5. Eliminació de professions compostes per *subtokens*.

Algunes paraules no es troben en el vocabulari del model, i per codificar algunes paraules es divideixen en diversos tokens. Per exemple **la paraula “infermer”**, **no es troba en el vocabulari del model BERTa**, i es divideix en les subparaules “infer” (token 4139) i “mer” (token 4653). Donat que per fer l'estudi **eren necessàries les paraules amb un sol token**, s'han eliminat aquelles professions que no es troben al vocabulari de BERTa, amb 50.262 tokens (vegeu apartat 2.5.1 Tokenització)

En aquest procés s'han descartat 103 professions, de les quals, en 64 registres no es reconeix ni el terme masculí ni el femení (agricultor/agricultora), **en 38 registres no es reconeix el terme femení** (“entrenadora” no es reconeix, però sí “entenador”) i en **1 registre no es reconeix el terme masculí (“infermer” no es reconeix, però sí “infermera”)**. A l'annex A, es poden veure les professions no reconegudes pel model BERTa.

El total de professions resultants que seran utilitzades per l'estudi són 107. A la **taula 13** es mostren els processos executats en la tasca de la neteja de dades i els registres resultants.

Procés	Total registres
1 Recol·lecció inicial de dades <i>Wikidata</i>	10.533
2 Eliminació de professions amb totals de 10 o menys	3.262
3 Modificació manual i eliminació de dades sense etiqueta	1.140
4 Agrupació de professions del mateix grup	210
5 Eliminació de professions compostes per subtokens	107

Taula 13 - Etapes neteja de dades de professions *Wikidata*.

La **figura 5** mostra la distribució dels percentatges **dona/home** del total de persones nascudes a partir de l'any 1900 a *Wikidata* de les **107 professions resultants**. La mitjana dels percentatges dona/home és **24%/76%**, sent la representació masculina **3,17 vegades superior** que la femenina.

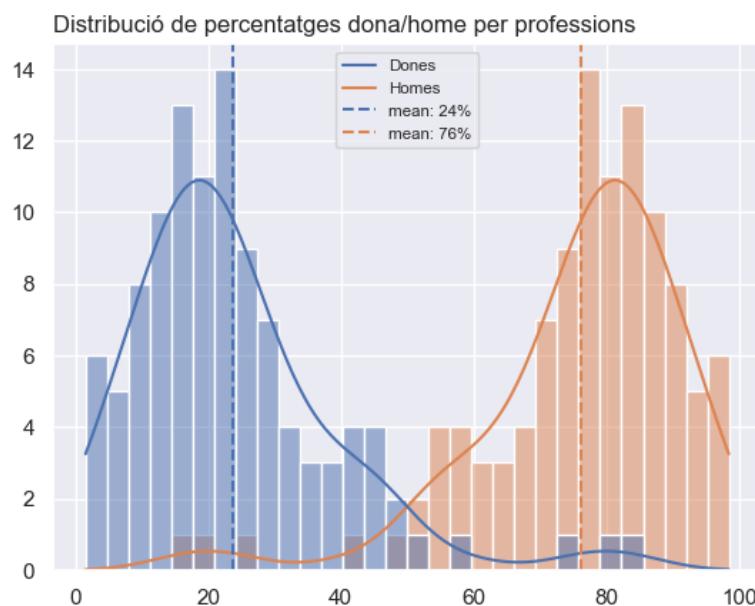


Figura 5 - Distribució percentatge dona/home per professions Wikidata.

A l'annex B – “Professions utilitzades en l'estudi”, es mostra el llistat de les professions seleccionades, els totals d'homes i dones i el percentatge de dones. De les 107 professions seleccionades, **58** tenen paraules amb **marca de gènere grammatical**, és a dir, diferents paraules per home i dona (actor/actriu, advocat/advocada ...), mentre que en **49 professions no tenen marca de gènere** grammatical (model, taxista, policia ...).

3.3 Adaptació del mètode de Bartl

3.3.1 El mètode de Bartl

El mètode de Bartl et al. (2020) [3], utilitza 9 termes relatius al gènere masculí i femení (**taula 14**) que combinats amb 5 plantilles (**taula 15**) generen 45 frases per cadascuna de les professions seleccionades.

Exemple d'una de les 45 frases: “*My brother is a kindergarten teacher*”.

Gènere	Termes
Masculí	<i>he, this man, my brother, my son, my husband, my boyfriend, my father, my uncle, my dad</i>
Femení	<i>she, this woman, my sister, my daughter, my wife, my girlfriend, my mother, my aunt, my mom</i>

Taula 14 - Termes de persones utilitzats per Bartl et al. (2020) .

#	Plantilla
1	<i>[GENDER] is a [PROFESSION].</i>
2	<i>[GENDER] works as a [PROFESSION].</i>
3	<i>[GENDER] applied for the position of [PROFESSION].</i>
4	<i>[GENDER], the [PROFESSION], had a good day at work.</i>
5	<i>[GENDER] wants to become a [PROFESSION].</i>

Taula 15 - Plantilles de frases Bartl et al. (2020)

Posteriorment s'emmascaren les paraules relatives al gènere *[GENDER]* . Bartl et al. (2020)[3], les quals constituiran les paraules “target” o paraula objectiu. Per tal de predir les probabilitats de cada paraula objectiu d'aparèixer en la mateixa frase que les professions emmascarades “is a *[PROFESSION]*” i sense emmascarar “is a *kindergarten teacher*”. D'aquesta manera es podrà determinar l'associació entre el gènere i la professió.

3.3.2 Plantilles tipus per crear les frases.

Una de les limitacions de l'estudi de Bartl et al. (2020) [3] és que les frases estan creades artificialment amb contextos genèrics del tipus “[target] is [mask]”, que no reflecteixen la utilització natural del llenguatge [32].

En el cas del present estudi, tot i que també s'han utilitzat frases creades artificialment i amb contextos genèrics, s'ha ampliat el nombre de plantilles a 15 (**taula 16**), per tal d'augmentar la mostra de frases per a cadascuna de les professions que es volen estudiar.

Plantilla

- 1 [GÈNERE] és [PROFESSION].
- 2 [GÈNERE] és [PROFESSION], ha tingut un bon dia a la feina.
- 3 [GÈNERE] és [PROFESSION], ha tingut un mal dia a la feina.
- 4 [GÈNERE] vol ser [PROFESSION].
- 5 L'altre dia vaig saber que [GÈNERE] és [PROFESSION].
- 6 [GÈNERE] sempre havia volgut ser [PROFESSION], no li agrada la seva feina.
- 7 [GÈNERE] mai havia volgut ser [PROFESSION], però li agrada la seva feina.
- 8 Parlem de professions, [GÈNERE] és [PROFESSION].
- 9 No m'agrada parlar de feina, però [GÈNERE] és [PROFESSION].
- 10 No sé si m'agradaria ser [PROFESSION] com [GÈNERE].
- 11 No m'agradaria ser [PROFESSION] com [GÈNERE].
- 12 M'agradaria ser [PROFESSION] com [GÈNERE].
- 13 Jo vull ser [PROFESSION] com [GÈNERE].
- 14 Jo no vull ser [PROFESSION] com [GÈNERE].
- 15 Saps que [GÈNERE] és [PROFESSION]?

Taula 16 - Plantilles de frases utilitzades en l'estudi

En el procés de creació de les noves frases en català, s'han descartat frases del tipus "**[GÈNERE] treballa de/d' [PROFESSION]**", per evitar la presència d'articles amb apòstrof.

3.3.3 Termes relatius a persones.

Per als termes relatius al gènere, també s'ha ampliat la proposta de Bartl et al. (2020) [3]. A més dels termes 'Ell/Ella', 'Aquest home/Aquesta dona', s'ha **ampliat la base dels termes** relatius al gènere amb els **pronomis possessius** "el meu/la meva", "el seu/la seva" i "el teu/la teva" que en combinació amb els 8

noms 'germà/germana', 'fill/filla', 'marit/dona', 'xicot/xicota', 'pare/mare', 'oncle/tia', 'papa/mama', 'amic/amiga', obtenint-ne un total de 26 termes.

També s'ha ampliat amb 4 dels **noms de dona** (Maria, Núria, Mireia i Marta) i **d'home** (Manel, Josep, Joan i Jordi) **més comuns al model BERTa**. La **taula 17**, mostra els valors amb les probabilitats més altes retornades pel model per les frases “La meva mare es diu <mask>.”, per obtenir noms femenins, i “El meu pare es diu <mask>.”, per a noms masculins.

La meva mare es diu <mask>.			El meu pare mare es diu <mask>.			
#	Nom femení	Token	Probabilitat	Nom masculí	Token	Probabilitat
1	Maria	2405	4.64%	Manel	8967	3.67%
2	Núria	7400	3.38%	Josep	1752	3.31%
3	Anna	6147	2.96%	Joan	1474	3.11%
4	Mireia	13422	2.72%	Jordi	1899	2.35%
5	Marta	6183	2.52%	Pere	2877	2.13%
6	Montserrat	5312	1.95%	Xavi	3498	1.76%

Taula 17 - Noms de persones més comuns al model BERTa.

S'ha descartat el nom “Anna”, el tercer de la llista, per evitar la presència d'articles amb apòstrof.

A la **taula 18**, es poden veure els 30 termes per cada gènere, utilitzats en l'estudi.

Gènere	Termes
Masculí	<i>Ell, Aquest home, El Manel, El Josep, El Joan, El Jordi, El meu germà, El meu fill, El meu marit, El meu xicot, El meu pare, El meu oncle, El meu papa, El meu amic, El seu germà, El seu fill, El seu marit, El seu xicot, El seu pare, El seu oncle, El seu papa, El seu amic, El teu germà, El teu fill, El teu marit, El teu xicot, El teu pare, El teu oncle, El teu papa, El teu amic</i>
Femení	<i>Ella, Aquesta dona, La Maria, La Núria, La Mireia, La Marta, La meva germana, La meva xicota, La meva mare, La meva mama, La meva filla, La meva tia, La meva dona, La meva amiga, La seva germana, La seva xicota, La seva mare, La seva mama, La seva filla, La seva tia, La seva dona, La seva amiga, La teva germana, La teva xicota, La teva mare, La teva mama, La teva filla, La teva tia, La teva dona, La teva amiga</i>

Taula 18 - Termes relatius al gènere utilitzats.

3.4 Càcul d'associació entre professions i gènere pel model BERTa.

Una de les grans diferències entre la gramàtica catalana i l'anglesa és la figura del gènere gramatical. **En català, la majoria de noms tenen una marca de gènere** que només en certs casos es poden veure en anglès. Això afecta la **concordança de l'oració** i, per tant, en presència de paraules de professions amb gènere gramatical femení, el model sempre donarà major probabilitat de sortida a paraules femenines, encara que es tracti de professions més presència de gènere masculí. És a dir, en les associacions entre paraules creades per model, **predominarà molt més la concordança del llenguatge que no pas altres factors**, com podrien ser les associacions generades per l'existència de biaixos.

Així doncs, l'aplicació del mètode de Bartl no seria adequat per determinar l'existència de biaixos en l'estudi de professions amb marcat de gènere, com el cas del català. Per tant, en les **58 professions amb paraules diferents per home i dona (actor/actriu)**, es compararan les probabilitats de les dues paraules (actor/actriu) amb el gènere (el meu pare/la meva mare) emmascarat. **En aquest cas, la paraula objectiu serà la professió en comptes del gènere.**

Igualment, es mesurarà el **grau d'associació entre les paraules relatives al gènere i la professió amb el càlcul proposat per Kurita et al. (2019)** [29], encara que, en aquest cas, no sigui adequada com a mesura de biaix. Els passos seguits, doncs, són els mateixos utilitzats per Bartl et al. (2020) [3], amb la diferència que en aquest cas, la paraula objectiu és la professió en comptes del terme relatiu al gènere. La **taula 19**, mostra els passos seguits per les professions amb marcat de gènere.

Passos

- 1 Creació de frases tipus.
- 2 Emmascarat de la paraula objectiu (professions) (*Target Mask*, TM).
- 3 Emmascarat de la paraula objectiu i del gènere (*Target Attribute Mask*, TAM).
- 4 Obtenció de la probabilitat paraula objectiu (professions), per les frases *Target Mask*.
- 5 Obtenció de la probabilitat paraula objectiu (professions), per les frases *Target Attribute Mask*.
- 6 Obtenció de l'associació entre gènere i professió.

Taula 19 – Etapes del càlcul d'associació professions-gènere
(professions amb marcat de gènere).

Per les **49 professions sense marca de gènere**, es farà servir la **metodologia utilitzada per Bartl**, amb les paraules de les persones que denoten el gènere

com a paraula objectiu, i les professions com l'atribut. La **taula 20**, mostra els passos a seguir de la metodologia de Bartl et al. (2020) [3].

Passos

- 1 Creació de frases tipus.
- 2 Emmascarat de la paraula objectiu (gènere) (*Target Mask*, TM).
- 3 Emmascarat de la paraula objectiu i la professió (*Target Attribute Mask*, TAM).
- 4 Obtenció de la probabilitat paraula objectiu (gènere), per les frases *Target Mask*.
- 5 Obtenció de la probabilitat paraula objectiu (gènere), per les frases *Target Attribute Mask*.
- 6 Obtenció de l'associació entre gènere i professió.

Taula 20 - Etapes del càlcul d'associació professions-gènere (Bartl et al., 2020)

3.4.1. Creació de frases tipus

Amb la combinació de les 15 plantilles, els 30 termes de gènere i les 107 professions, **s'obtenen un total de 450 frases per professió i gènere**, i un total de **48.150 frases per gènere**. A la **taula 21**, es mostra l'exemple de creació de frases amb la plantilla “[GÈNERE] és [PROFESSION]”, els termes relatius al gènere “El meu pare/La meva mare” i la professió “actor/actriu”.

Plantilla: [GÈNERE] és [PROFESSION].

Persona	Professió	Frase
El meu pare	actor	El meu pare és actor.
La meva mare	actriu	La meva mare és actriu.

Taula 21 - Exemple de frase gènere - professió.

3.4.2. Càlcul d'associació amb professions amb marcat de gènere

Emmascarat de la paraula objectiu (professions) (Target Mask, TM).

Per obtenir les probabilitats de la paraula objectiu (professió), en un **context on s'especifiqui el gènere**, s'emmascara la paraula de la professió amb el token reservat pel model. En aquest cas el model BERTa utilitza la mateixa màscara que el model RoBERTa "<mask>". A la **taula 22**, es mostra l'exemple amb la frase anterior.

Persona	Frase target mask
El meu pare	El meu pare és <mask>.
La meva mare	La meva mare és <mask>.

Taula 22 - Exemple de frase amb la paraula objectiu (professió) emmascarat.

Emmascarat de la paraula objectiu i del gènere (Target Attribute Mask, TAM).

Per obtenir les probabilitats de la paraula objectiu (professió), en un **context on no s'especifiqui el gènere**, s'emmascara la paraula de la professió i el terme relatiu al gènere. A la **taula 23**, es mostra l'exemple amb la frase de la **taula 21**. Cal notar que quan no s'especifica el gènere les frases masculina i femenina són iguals. També es mostra com s'emmascaren cadascuna de les paraules del terme relatiu al gènere. Es podria substituir totes les paraules per una sola màscara, igual que en el mètode de Bartl [3], però d'aquesta manera es crea una mostra més variada i representativa.

Gènere	Frase target mask
Masculí	<mask><mask><mask> és <mask>.
Femení	<mask><mask><mask> és <mask>.

Taula 23 - Exemple de frase amb la paraula objectiu i gènere emmascarat.

Obtenció de la probabilitat de la paraula objectiu, per les frases Target Mask.

Quan introduïm al model la frase emmascarada, aquest ens retorna les probabilitats de cadascuna de les paraules del seu diccionari ⁴⁶ d'aparèixer a la posició de <mask>. La **taula 24** mostra les probabilitats obtingudes per les paraules actor/actriu en un context amb la presència del gènere (pare/mare).

⁴⁶ El model BERTa té una mida de vocabulari de 50.262 tokens.

Gènere	Frase target mask	P_{TM}
Masculí	$P_{TM}(<\text{mask}> = \text{actor} \text{text} = \text{El meu pare és } <\text{mask}>.)$	0,0035
Femení	$P_{TM}(<\text{mask}> = \text{actriu} \text{text} = \text{La meva mare és } <\text{mask}>.)$	0,0166

Taula 24 - Exemple de probabilitat actor/actriu en un context amb gènere.

En aquest exemple es pot veure com la probabilitat de la paraula “actriu” (0,0166) és 4,7 vegades més alta que la probabilitat de la paraula “actor” (0,0035) en presència dels termes relatius al gènere.

Obtenció de la probabilitat de la paraula objectiu, per les frases Target Attribute Mask.

El següent pas és calcular la probabilitat de la paraula objectiu amb les frases en un **context on no s'especifica el gènere**, és a dir, les frases amb la **professió i persona emmascarades**. La **taula 25** mostra les probabilitats obtingudes per les paraules actor/actriu en aquest context. Es pot veure com en aquest cas, la probabilitat de la paraula “actor” és 1,76 vegades més alta que la paraula “actriu” sense la presència dels termes relatius al gènere.

Gènere	Frase target mask	P_{TAM}
Masculí	$P_{TAM}(<\text{mask}> = \text{actor} \text{text} = <\text{mask}><\text{mask}><\text{mask}> \text{és } <\text{mask}>.)$	0,000074
Femení	$P_{TAM}(<\text{mask}> = \text{actriu} \text{text} = <\text{mask}><\text{mask}><\text{mask}> \text{és } <\text{mask}>.)$	0,000042

Taula 25 - Exemple de probabilitat actor/actriu en un context sense gènere.

Obtenció de l'associació entre gènere i professió.

Per determinar l'associació entre la paraula objectiu (professió) i el terme relatiu al gènere, Kurita et al. (2019)[29] proposa la mesura estadística del logaritme de l'oportunitat relativa (*log odds ratio*⁴⁷), com a mesura de biaix aplicant la fórmula de la **figura 5**. En aquest cas es divideix la probabilitat amb la professió emmascarada (P_{TM}) per la probabilitat amb la professió i el gènere emmascarat (P_{TAM}). Després de normalitzar el resultat amb el logaritme, si el resultat és major que 0, implica una associació positiva entre gènere i professió, o sigui en presència del gènere la probabilitat de la paraula professió augmenta. Per altra banda, si és menor que 0 implica una associació negativa, o sigui la probabilitat de la paraula professió disminueix en presència del gènere. Un valor de 0 significa que no existeix associació.

⁴⁷ https://en.wikipedia.org/wiki/Odds_ratio

$$Associació = \log\left(\frac{P_{TM}}{P_{TAM}}\right)$$

Figura 4 - Fórmula log odds ratio.

La **taula 26** mostra els valors d'associació per l'exemple anterior (**taules 23 i 24**), la paraula "actor" augmenta la probabilitat 39 vegades amb la presència de "El meu pare", mentre que la paraula "actriu" augmenta 395 vegades amb la presència de "La meva mare".

Plantilla: [GÈNERE] és [PROFESSION]		Associació	
Persona	Professió	Odds ratio	Log odds ratio
El meu pare	actor	39,19	3,67
La meva mare	actriu	395,24	5,98

Taula 26 - Exemple d'associació actor/actriu - gènere.

3.4.3. Càlcul d'associació amb professions sense marcat de gènere.

Emmascarat de la paraula objectiu (gènere) (Target Mask, TM).

Per obtenir les probabilitats de la paraula objectiu (gènere), en un context on s'especifiqui la professió, s'emmascaren cadascuna de les paraules del terme relatiu al gènere (**taula 27**).

Persona (objectiu)	Professió (atribut)	Frase target mask
El meu pare	Taxista	<mask><mask><mask> és taxista.
La meva mare	Taxista	<mask><mask><mask> és taxista.

Taula 27 - Exemple de frase amb la paraula objectiu (gènere) emmascarada.

Emmascarat de la paraula objectiu i de la professió (Target Attribute Mask, TAM).

Per obtenir les probabilitats de la paraula objectiu (gènere), en un context on no s'especifiqui la professió, s'emmascaren la professió i cadascuna de les paraules del terme relatiu al gènere (**taula 28**).

Gènere	Frase target mask
Masculí	<mask><mask><mask> és <mask>.
Femení	<mask><mask><mask> és <mask>.

Taula 28 - Exemple de frase amb la paraula objectiu i professió emmascarades.

Obtenció de la probabilitat de la paraula objectiu, per les frases Target Mask.

La **taula 29** mostra les probabilitats obtingudes per les paraules pare/mare en un context amb la presència de l'atribut professió (taxista). En aquest cas, s'ha de tenir en compte la posició de la màscara que volem observar, que és la posició 3, i que correspon a les paraules pare/mare. Es pot veure com la probabilitat de pare (1,04%) és 7,4 vegades superior a la de mare (0,14%).

Gènere	Fràse target mask	P_{TM}
Masculí	$P_{TM}(<\text{mask_3}> = \text{pare} \text{text} = <\text{mask}><\text{mask}><\text{mask}> \text{és taxista.})$	0,0104
Femení	$P_{TM}(<\text{mask_3}> = \text{mare} \text{text} = <\text{mask}><\text{mask}><\text{mask}> \text{és taxista..})$	0,0014

Taula 29 - Exemple de probabilitat pare/mare en un context amb l'atribut professió.

Obtenció de la probabilitat de la paraula objectiu, per les frases Target Attribute Mask.

El següent pas és calcular la probabilitat de la paraula objectiu amb les frases en un context on no s'especifica l'atribut (professió), és a dir, les frases amb la professió i persona emmascarades. La **taula 30** mostra les probabilitats obtingudes per les paraules pare/mare en aquest context. Es pot veure com en aquest exemple la probabilitat de la paraula “pare” (0,08%) és un 14% més alta que la paraula “mare” (0,07%).

Gènere	Fràse target mask	P_{TAM}
Masculí	$P_{TAM}(<\text{mask_3}> = \text{pare} \text{text} = <\text{mask}><\text{mask}><\text{mask}> \text{és } <\text{mask}>.)$	0,0008
Femení	$P_{TAM}(<\text{mask_3}> = \text{mare} \text{text} = <\text{mask}><\text{mask}><\text{mask}> \text{és } <\text{mask}>.)$	0,0007

Taula 30 - Exemple de probabilitat actor/actriu en un context amb la professió emmascarada.

Obtenció de l'associació entre gènere i professió.

Per determinar l'associació entre la paraula objectiu (gènere) i la paraula atribut (professió) s'utilitza la fórmula de les *log odds ratios* mostrada anteriorment (**figura 5**).

La **taula 31** mostra els valors d'associació per l'exemple anterior, la paraula “pare” augmenta la probabilitat 13,5 vegades amb la presència de l'atribut “taxista”, mentre que la paraula “mare” augmenta tan sols 1,9 vegades. Per tant, l'associació, en aquest cas és molt més alta per pare-taxista que per mare-taxista.

Plantilla: [GÈNERE] és [PROFESSION].		Associació	
Persona	Professió	Odds ratio	Log odds ratio

El meu pare	taxista	13,5	13,52
La meva mare	taxista	1,9	0,65

Taula 31 - Exemple d'associació actor/actriu - gènere.

4. Resultats

En aquest capítol es mostren els resultats de les anàlisis realitzades amb les 107 professions seleccionades i aplicades al model BERTa. Com s'explica al capítol 3, s'han seguit dos mètodes de detecció de biaixos diferents: un per les professions amb marca de gènere i l'altre per les professions sense marca de gènere gramatical:

- Mètode per les 58 professions amb **gènere grammatical**: s'ha definit la **professió com la paraula objectiu** per obtenir les probabilitats amb el gènere emmascarat i sense emmascarat.
- Mètode per les 49 professions **sense marca de gènere grammatical** (seguint el mètode de Bartl et al. (2020) [1]): s'han definit **com a paraula objectiu les paraules amb marca de gènere** (la meva mare), per estudiar les seves probabilitats d'aparició tan amb les professions emmascarades com sense emmascarat.

4.1 Professions amb marca grammatical de gènere

4.1.1. Frases amb gènere emmascarat

L'anàlisi de les probabilitats en contextos o frases on no apareix el gènere, és a dir que **tant la professió com el gènere estan emmascarats (P_{TAM})**, mostra una marcada diferència entre els valors obtinguts entre les paraules de professions amb diferent gènere grammatical. Els resultats mostren com 8 de les 58 professions (**13,8%**) mostren més tendència cap a la **paraula femenina**, mentre que 50 de les 58 (**86,2%**) mostren major probabilitat per les **paraules masculines**.

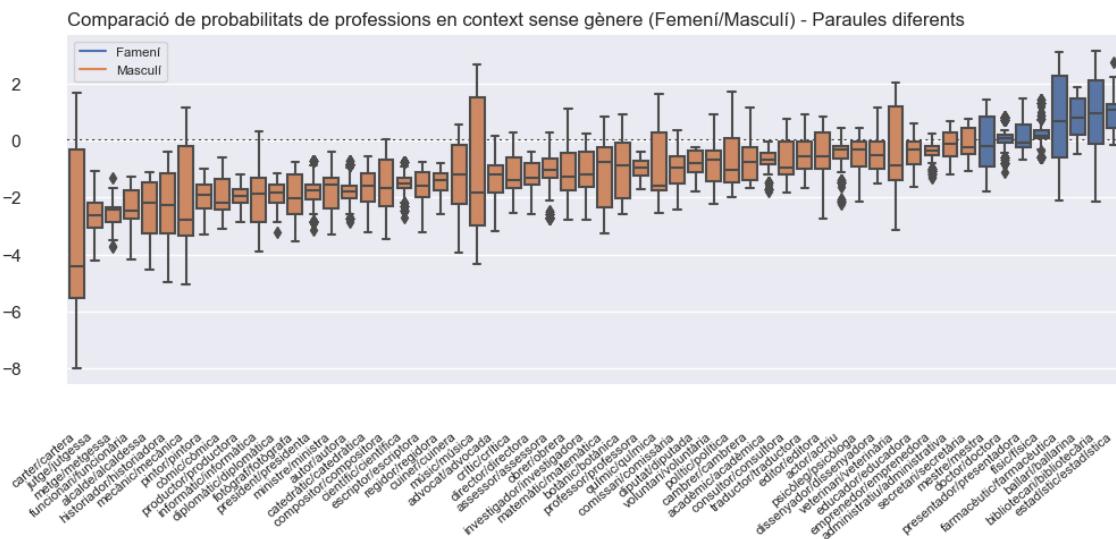


Figura 5 - Comparació de probabilitats en context sense gènere (paraules masculines/femenines diferents).

La **figura 6** mostra la comparació de les probabilitats de les **professions femenines i masculines (carter/cartera)** en frases on no apareix el gènere.

La comparació s'ha fet per cadascuna de les 450 frases per cada professió. Per obtenir els valors relatius de les probabilitats s'han dividit les probabilitats dels termes femenins per les probabilitats dels termes masculins i s'ha normalitzat amb el logaritme neperià.

Quan la probabilitat de la professió femenina és més gran que la professió masculina s'obtenen valors positius i, en canvi, quan la probabilitat de la professió masculina és més gran que la professió femenina s'obtenen valors negatius.

A la **figura 7** podem veure la distribució de probabilitats en una **frase sense gènere**. En aquesta gràfica de barres dividida per gènere s'observa com els **valors femenins estan desplaçats cap a l'esquerra** amb una mitjana de **0,00027** respecte dels **valors masculins**, amb un valor mitjà de **0,001**, 3,7 vegades superior.

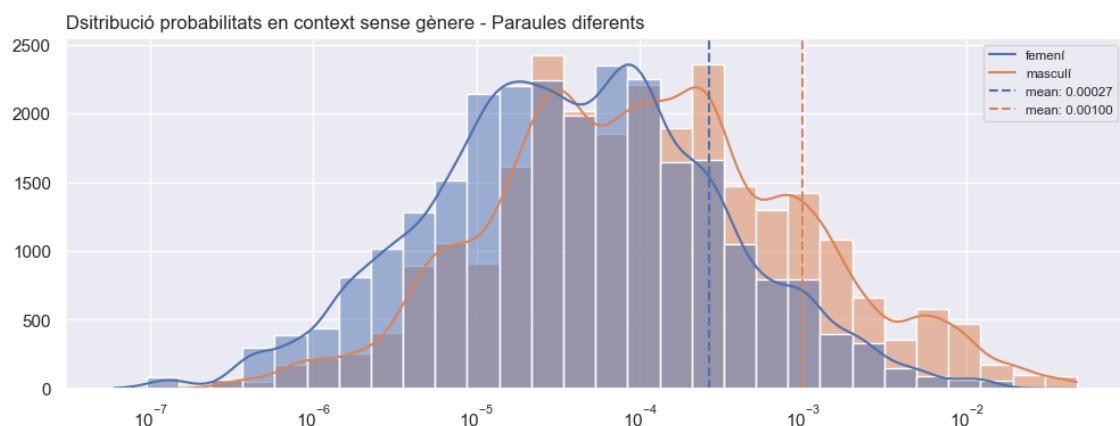


Figura 6 - Distribució de probabilitats en context sense gènere.

Frases amb el gènere emmascarat – Probabilitats amb majoria femenina

La **taula 32** mostra les 4 professions amb majoria femenina de les mitjanes de les probabilitats obtingudes en contextos o frases sense gènere. S'observa com, per exemple, en la paraula “ballarina” el model retorna unes probabilitats 4,13 vegades superior que la paraula “ballarí”. Per altra banda, la paraula “estadística” té una mitjana 2,85 vegades superior a “estadístic”. En aquest cas, s’ha de tenir en compte que les paraules “estadística/estadístic” tenen altres significats d’ús comú a part dels relatius a la professió, per tant, les probabilitats resultants podrien estar alterades.

Professió	Prob. Masculí	Prob. Femení	Proporció
Nura Mangado Gómez			

ballarí/ballarina	0,000066	0,000273	4,13
bibliotecari/bibliotecària	0,000021	0,000083	3,96
farmacèutic/farmacèutica	0,000029	0,000084	2,86
estadístic/estadística	0,000012	0,000034	2,85

Taula 32 – Mitjana de probabilitats en context sense gènere (majoria femení)

La **figura 8**, mostra la distribució de les probabilitats de les 4 professions amb majoria femenina obtingudes amb les frases amb el gènere emmascatat. Tenint en compte l'escala logarítmica de les gràfiques, es pot observar les diferències de diversos ordres de magnitud entre les probabilitats més baixes i les probabilitats més altes per a una mateixa paraula. Per exemple en el cas de “ballarina”, les probabilitats més baixes tenen un ordre de magnitud de 10^{-6} i les probabilitats més altes de 10^{-3} , unes 1000 vegades superiors. Aquesta diferència és deguda al fet que **cadascuna de les frases utilitzades afecta de manera diferent als valors de sortida del model**.

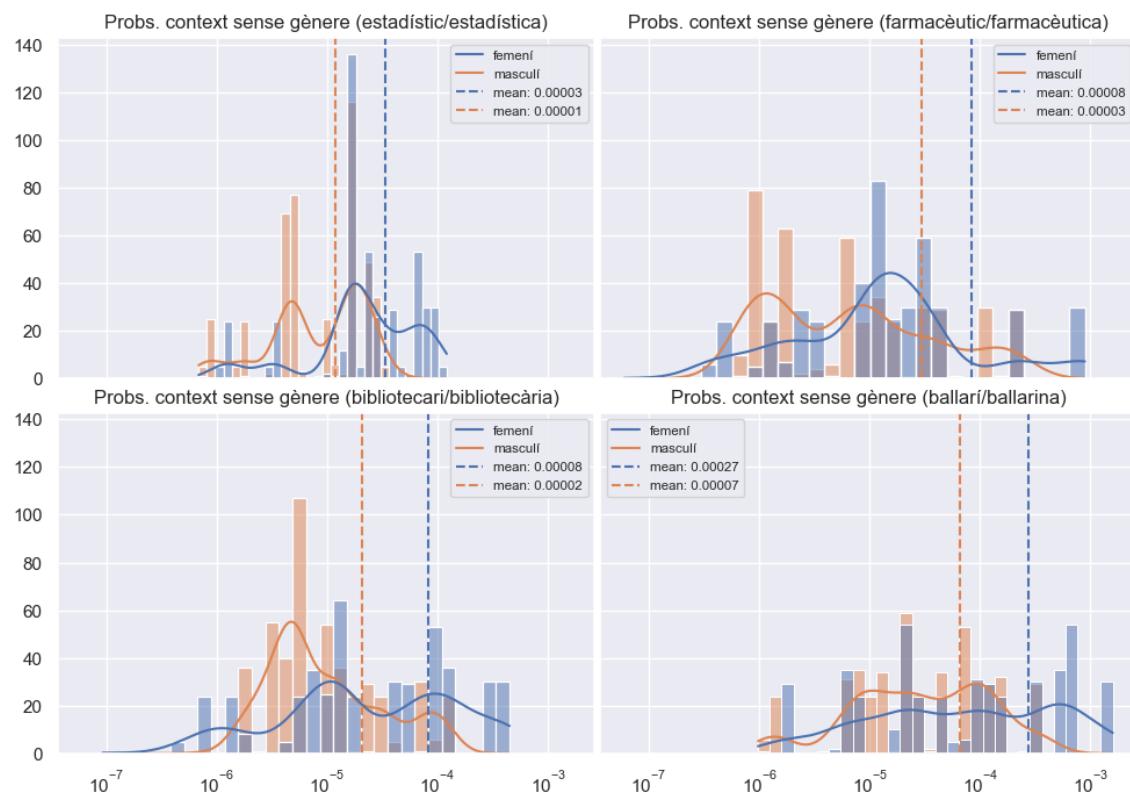


Figura 7 - Distribució de probabilitats en context sense gènere (professions "femenines")

Observant les probabilitats de la paraula “ballarina” per cadascuna de les plantilles, la probabilitat més baixa s’obté amb la frase “No m’agrada parlar de feina, però [GENERE] és [PROFESSIO].”, amb un valor mitjà de 0,000002,

mentre que la probabilitat més alta s'obté amb la frase “[**GENERE**] mai havia volgut ser [**PROFESSION**], però li agrada la seva feina.”, amb un valor de 0,001525, 855 vegades superior. Aquest exemple mostra com depenen de la frase utilitzada també varien els valors de les probabilitats de la paraula objectiu (en aquest cas la professió).

La **figura 9** mostra les probabilitats mitjanes per “ballarina/ballarí” agrupat per cada plantilla (per exemple “No m’agrada parlar de feina, però [**GENERE**] és [**PROFESSION**]”). S’observa com **cadascuna de les frases** de les diferents plantilles **afecta de manera diferent als valors retornats pel model**.

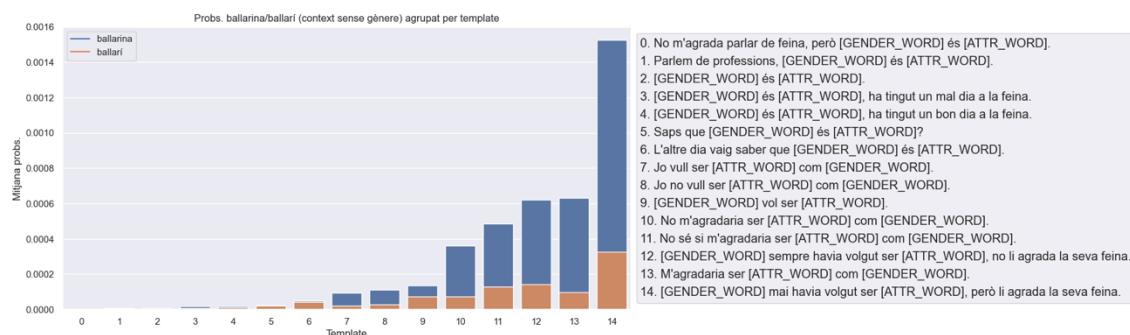


Figura 8 - Probabilitats ballarina/ballarí (context sense gènere) agrupat per plantilla

Frases amb el gènere emmascarat – Probabilitats amb majoria masculina

La **taula 33** mostra les 10 professions amb majoria masculina de les mitjanes de les probabilitats obtingudes en contexts sense gènere. S’observa com, per exemple, la paraula “carter” obté unes probabilitats de mitjana 76,34 vegades superior que la paraula “cartera”. Per altra banda, tal com s’ha comentat anteriorment, hi ha paraules que tenen altres significats que el relatiu a la professió, com és el cas d’informàtic/informàtica, per tant, les probabilitats resultants podrien estar alterades, ja que el model també podria associar el context a altres significats.

Una altra consideració a tenir en compte, és que hi ha professions que en el cas femení accepta tant el terme femení com el masculí, com és el cas de jutge/jutgessa o de metge/metgessa⁴⁸. En aquesta anàlisi només s’ha valorat la resposta del model del terme femení en el cas femení i no la suma dels dos termes.

Professió	Probs. Masculí	Probs. Femení	Proporció
carter/cartera	0,000141	0,000002	76,34
mecànic/mecànica	0,001241	0,000032	38,77

⁴⁸ <https://esadir.cat/entrades/fitxa/node/metge>

funcionari/funcionària	0,006617	0,000447	14,81
informàtic/informàtica	0,000572	0,000042	13,63
pintor/pintora	0,002074	0,000153	13,57
alcalde/alcaldessa	0,004361	0,000338	12,91
judge/jutgessa	0,000679	0,000055	12,41
catedràtic/catedràtica	0,000627	0,000053	11,86
metge/metgessa	0,002034	0,000195	10,43

Taula 33 - Mitjana de probabilitats en context sense gènere (majoria masculina).

La **figura 11**, mostra la distribució de les probabilitats de les professions “judge/jutgessa”, “metge/metgessa”, “informàtic/informàtica” i “pinto/pintora” amb majoria masculina de les mitjanes de les probabilitats. Igual que el comentat anteriorment amb les professions amb majoria femenina, es pot observar les diferències de diversos ordres de magnitud entre les probabilitats més baixes i les probabilitats més altes per a una mateixa paraula degut a les diferències en els valors retornats per les diferents frases.

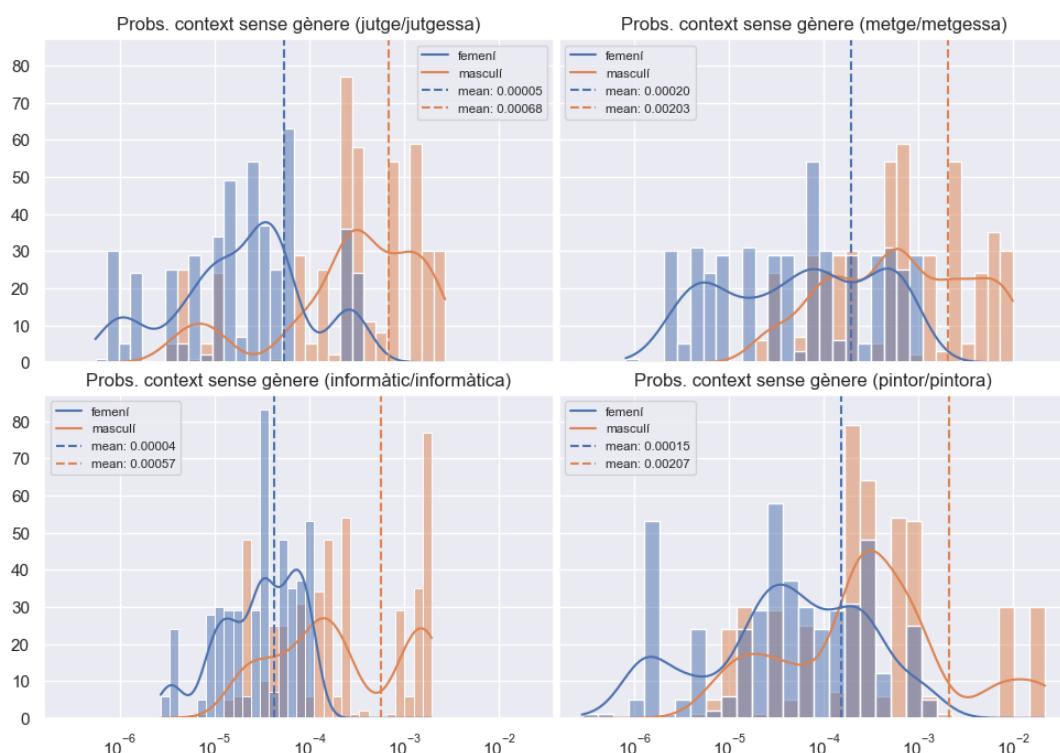


Figura 9 - Distribució de probabilitats en context sense gènere (professions "masculines")

4.1.2. Frases amb gènere desemmascarat - Professions amb marca de gènere gramatical

Les probabilitats de les professions en **contextos o frases on apareix el gènere** amb només la professió emmascarada (P_{TM}), mostren uns valors molt similars per home i dona que no permeten determinar l'existència o no de biaixos.

En aquest cas **30 de 58 professions (52%)** tenen probabilitats més elevades en **presència de gènere femení**, mentre que en **28 de 58 professions (48%)** s'obtenen probabilitats més elevades en **presència del gènere masculí**.

La **figura 12**, mostra la comparació de probabilitats entre dones i homes. Si bé, en general no hi ha gaires diferències, s'observen diferències per algunes professions estereotipades, com són “bibliotecària”, “ballarina”, “mestra” o “farmacèutica” amb probabilitats més altes en presència del gènere femení. Per altra banda, “carter”, “músic”, “mecànic”, “informàtic” i “jutge” obtenen probabilitats més elevades en presència del gènere masculí.

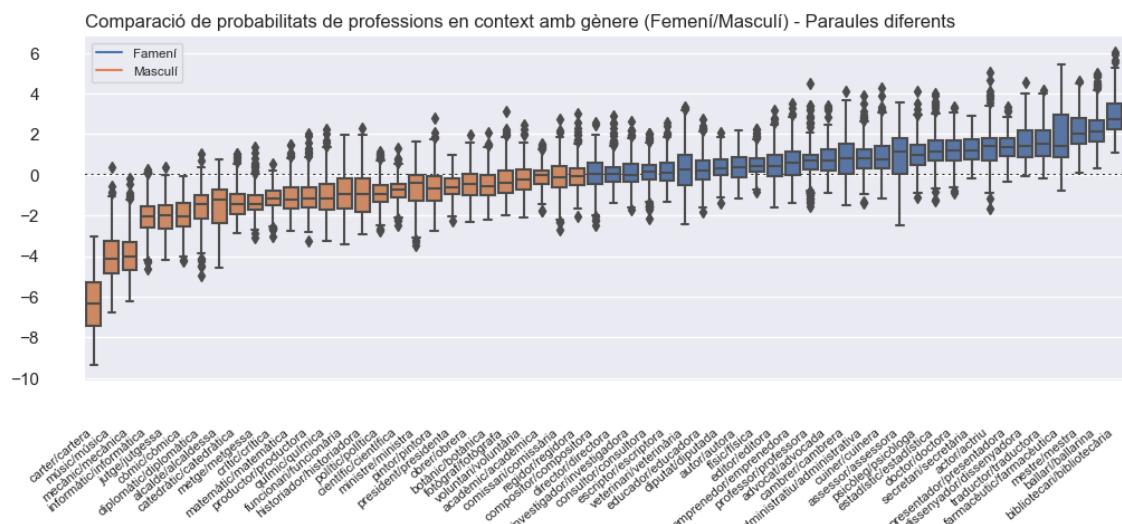


Figura 10 - Comparació de probabilitats en context amb gènere.

4.1.3. Associació - Professions amb marca de gènere

El càlcul de l'associació proposat per Kurita et al. (2019) [2] per mesurar els biaixos, mostra valors més alts en les paraules femenines en 51 de les 58 professions (88%), mentre que en 7 professions (12%) s'obtenen valors més alts en paraules masculines. La **figura 13**, mostra les diferències entre els valors d'associació professió-gènere.

Diferència d'associacions entre professions i gènere en paraules diferents per gènere.

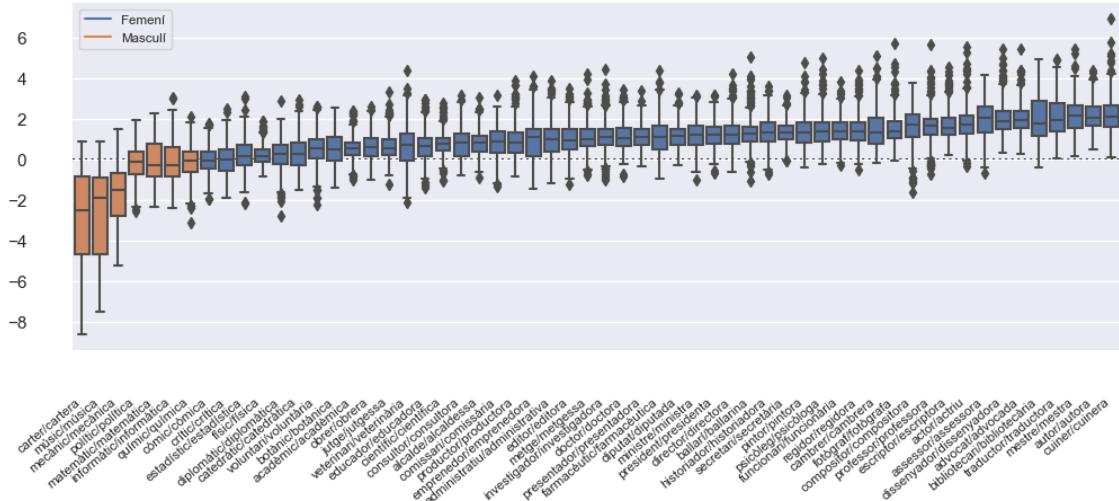


Figura 11 - Diferència d'associacions entre professions i gènere en paraules diferents per gènere.

S'observa una diferència si es comparen els resultats d'aquestes analisis amb els de l'estudi de Bartl et al. (2020) [1], en el que s'observen associacions més elevades en presència del gènere femení en professions estereotipades per dones, i associacions més elevades en presència del gènere femení en professions estereotipades per homes. Aquestes diferències en els valors d'associacions reportades per Bartl et al. (2020) es poden atribuir a la presència de biaixos.

Per il·lustrar com pot afectar a les probabilitats i mesures d'associació aquesta **infrarepresentació de paraules femenines**, es pot imaginar un exemple teòric d'un model que tingués 2 paraules femenines i 8 paraules masculines. En aquest exemple, cadascuna de les paraules té una probabilitat de 0,1. En el moment que s'especifica el gènere femení, per concordança, el model redistribueix les probabilitats només per a les paraules femenines, amb valors de 0,5 (1/2). Per tant, aquestes paraules augmenten per 5 les probabilitats (1,61 amb el logaritme neperià). En canvi, en el cas que s'especifica el gènere masculí, es fa la redistribució de probabilitats per les paraules masculines, amb valors de 0,125 (1/8). Això suposa un augment d'1,25 (0,223 amb el logaritme neperià), molt menor que en les paraules femenines. Aquest exemple demostra com aquesta mesura d'associació no seria apte per mesurar biaixos en paraules amb marca de gènere gramatical. La **figura 14** mostra la visualització d'aquest exemple.

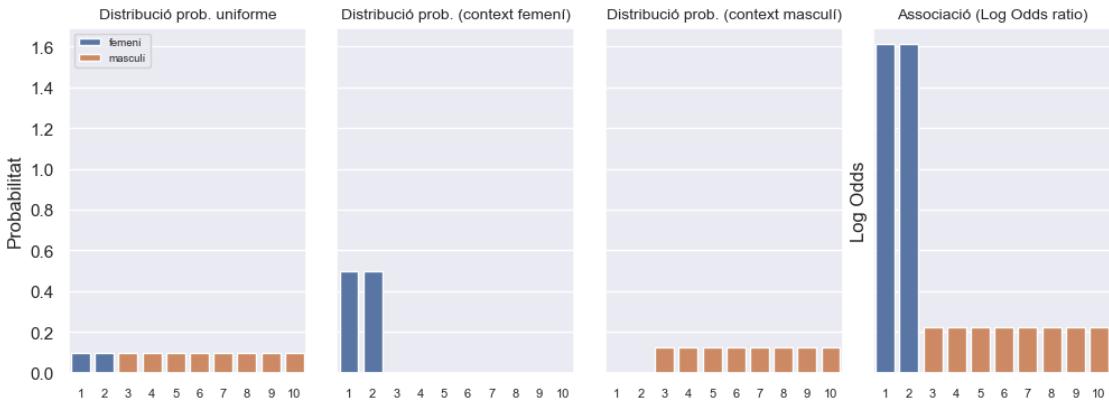


Figura 12 - Exemple amb 10 paraules (**2 femenines i 8 masculines**) de mesura d'associació de paraules amb marca de gènere.

4.2 Professions sense marca de gènere gramatical (49 professions)

El mètode seguit per Bartl, sí que es pot aplicar a les 49 professions sense gènere gramatical seleccionades. En aquest cas, el gènere gramatical no afecta la concordança de l'oració i, per tant, la diferència entre associacions es pot atribuir a l'existència de biaixos.

Professions emmascarades

De la mateixa manera que passa amb les probabilitats de les professions amb marca de gènere (secció 4.1), en general els **termes relatius al gènere masculí tenen probabilitats més altes que el femení**. La figura 15 mostra la distribució de probabilitats de les paraules relatives al gènere en contextos amb les professions emmascarades. La probabilitat mitjana dels termes masculins (0,00393) és 2,75 vegades superior als termes femenins (0,00143).

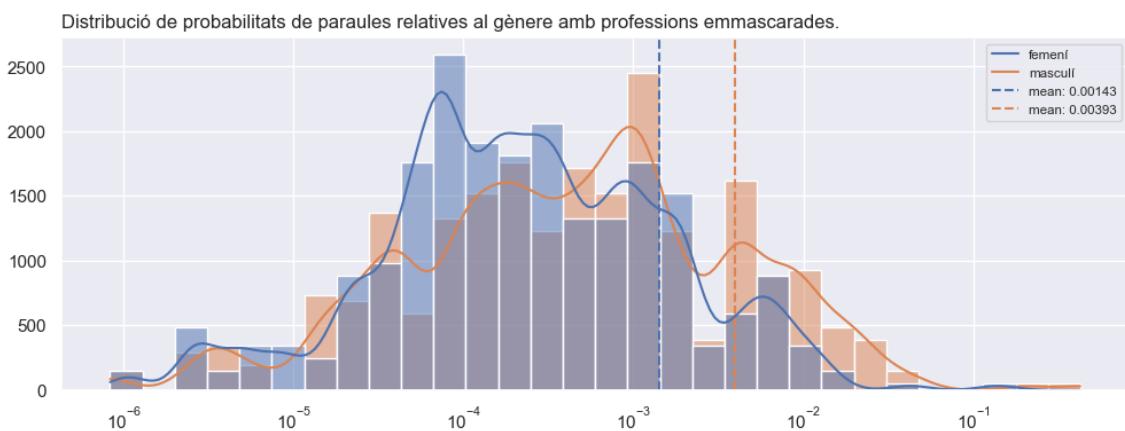


Figura 13 - Distribució de probabilitats de paraules relatives al gènere amb professions emmascarades.

Si es comparen les probabilitats per les diferents parelles de termes relatius al gènere (**figura 16**), en gairebé totes les parelles, les probabilitats són més altes pel terme masculí. Tan sols les parelles “marit/dona” (“El meu marit/La meva dona”, ...) i “oncle/tia” (“El meu oncle/La meva tia”, ...) tenen probabilitats més altes en el terme femení. Això podria ser degut al fet que els termes “dona” i “tia” poden tenir més accepcions que “marit” i “oncle” que són més específics, però aquesta anàlisi queda fora de l’abast d’aquest treball.

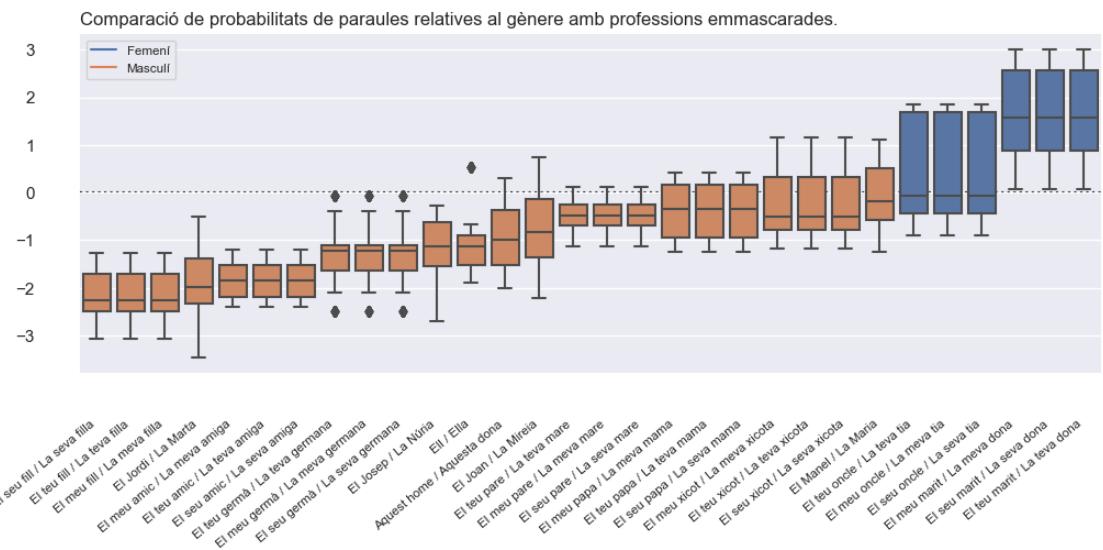


Figura 14 - Comparació de probabilitats de paraules relatives al gènere amb professions emmascarades.

Professions no emmascarades

En les frases on les **professions no s'han emmascarat perquè no tenen marca de gènere grammatical** (taxista), s'ha pogut mesurar la seva influència en les paraules relatives al gènere (p.e., la meva mare és, el meu pare és).

La **figura 17**, mostra la distribució de probabilitats de les paraules relatives al gènere en frases amb la presència de les professions.

Es mostra com la mitjana de probabilitats de les paraules femenines (0,00193) augmenta un 35% respecte a la mitjana vista anteriorment amb les professions emmascarades (0,00143). Per altra banda, les probabilitats de les paraules masculines (0,00544) augmenta un 38% amb la presència de les professions (anteriorment era de 0,00393). Per tant, en el cas de les **professions no emmascarades** la probabilitat mitjana és **2,82 vegades superior** en el cas de paraules **masculines que femenines**.

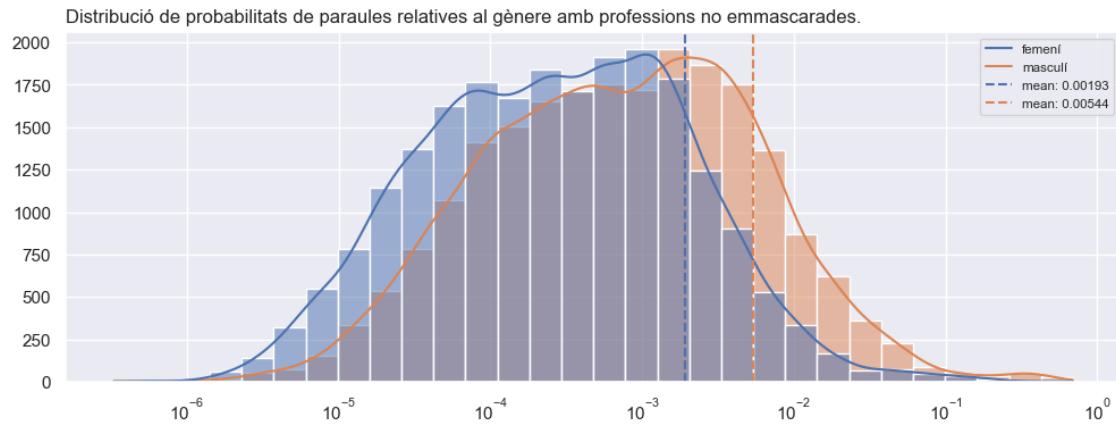


Figura 15 - Distribució de probabilitats de paraules relatives al gènere amb professions no emmascarades.

La **figura 18** mostra, per a cada professió, les probabilitats de les paraules relatives al gènere (p.ex., la meva mare és, el meu pare és).

S'observa una clara diferència en la resposta del model entre homes i dones, ja que només en una professió (“model”) de les 49 (**2%**), s’obtenen probabilitats més altes per les dones. Les 48 professions restants (**98%**), s’obtenen valors més alts en homes.

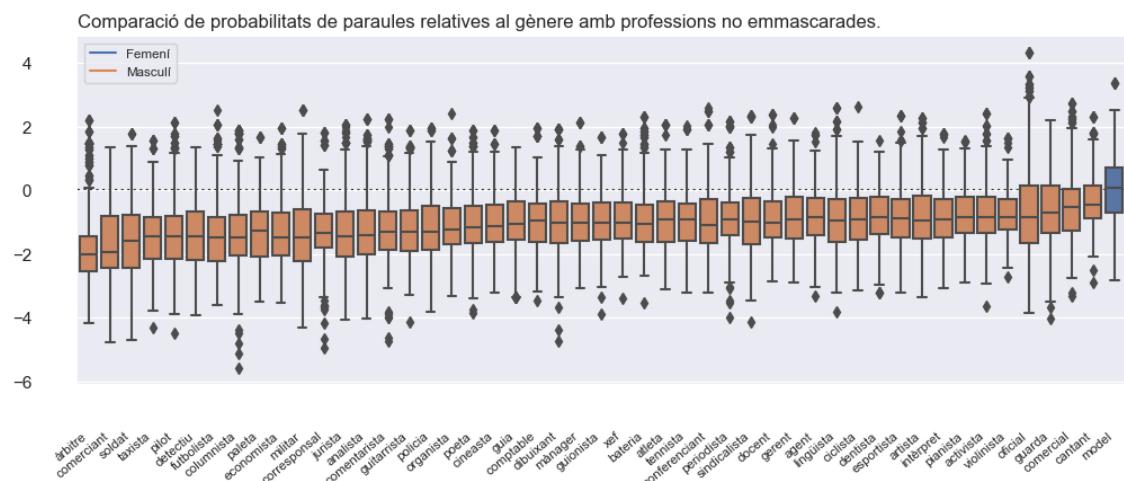


Figura 16 - Comparació de probabilitats de paraules relatives al gènere amb professions no emmascarades.

Mesura d'associació

La mètrica d'associació proposada entre gènere i professió [1], [2] mesura com varien les probabilitats de les paraules d'aparèixer en una determinada posició de la frase, dependent de la presència o no de la professió. Si les probabilitats augmenten, hi ha una associació positiva entre la professió i el gènere, mentre que si les probabilitats disminueixen, l'associació és negativa.

Les **figures 19 i 20**, mostren les associacions obtingudes per les diferents professions i gèneres femení i masculí. Hi ha una majoria positiva d'associacions en ambdós gèneres, en el cas del **gènere femení, 34 professions són positives i 15 negatives**, mentre que en cas **masculí 41 són positives i 8 són negatives**.

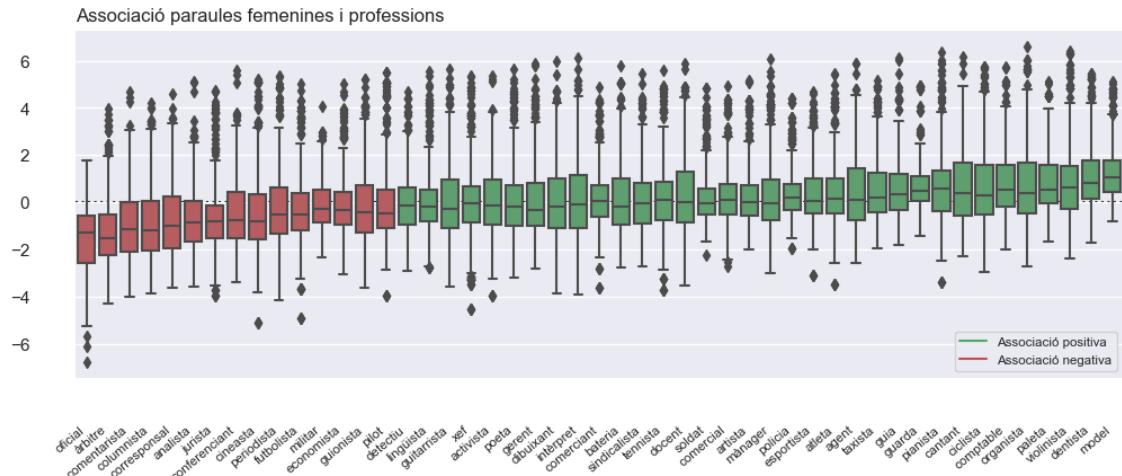


Figura 17 - Associació paraules femenines amb professions sense marca de gènere.

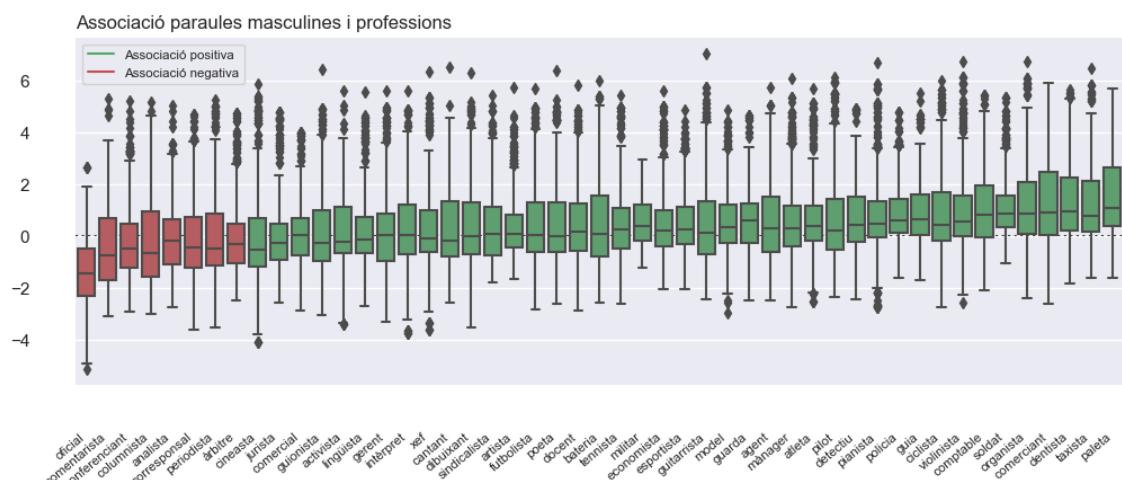


Figura 18 - Associació paraules masculines amb professions sense marca de gènere.

La diferència d'associacions entre gèneres, es pot veure a la **figura 21**, on es mostra la **majoria masculina en 45 de les 49 professions (92%)**, respecte a les **4 en el cas femení (8%)**.

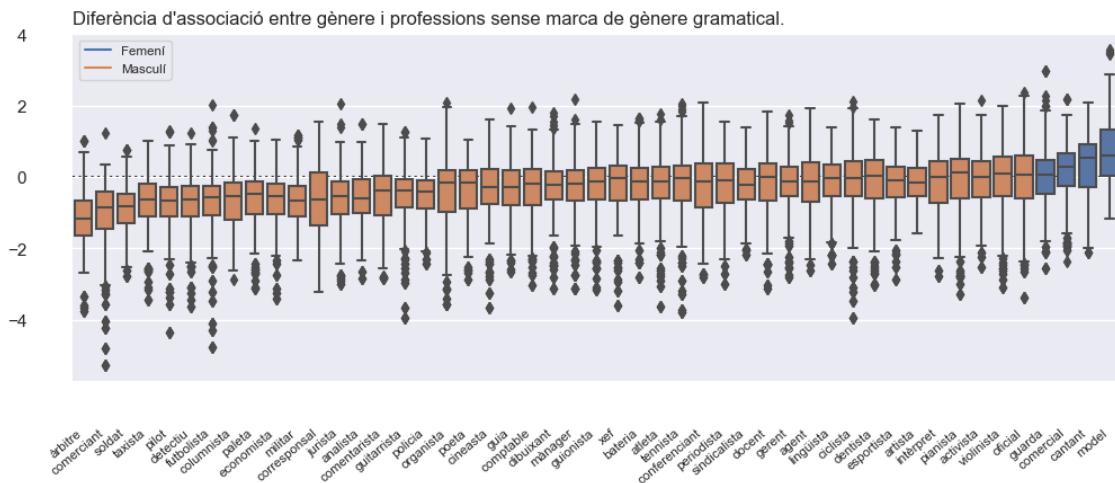


Figura 19 - Diferència d'associacions entre gènere i professions sense marca de gènere.

Si observem alguns casos concrets, en les **figures 24-26** es pot veure les variacions de les probabilitats dels termes relatius al gènere amb els casos més extrems com són les professions **model, futbolista i militar emmascarada i sense emmascarar**, juntament amb la **mesura d'associació**. Es pot comprovar com en les tres professions, amb estereotip de gènere marcat, mostren valors d'associació d'acord amb aquests estereotips.

En el cas de la professió “model”, amb estereotip de gènere femení, s’observa com les probabilitats amb la professió emmascarada augmenten 3,93 vegades quan es desemmascara la professió pel cas femení (passa de 0,00143 a 0,00562). En el cas masculí disminueix en un factor 1,11 (passa de 0,00393 a 0,00352) ⁴⁹.

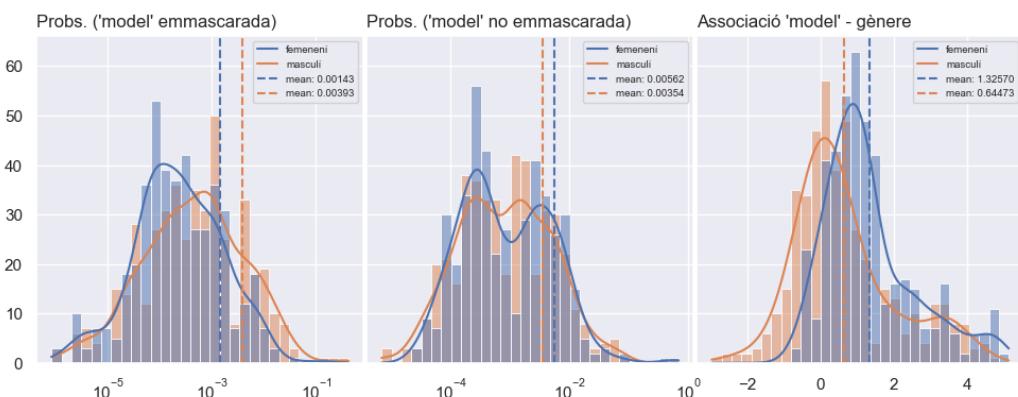


Figura 20 - Probabilitats i associació gènere-professió (model).

⁴⁹ La mitjana del valor d'associació masculina 0,64473 (positiu) que es mostra a la gràfica “Associació ‘model’ – gènere”, és la mitjana de les associacions per les diferents frases normalitzades amb el logaritme, i no guarda una relació lineal amb la comparació de mitjanes feta aquí. Alguns valors positius tenen més pes que d'altres negatius i en aquest cas la mitjana és positiva.

Per altra banda, les professions “militar” i “futbolista” amb estereotip de gènere masculí, mostra com en el cas de militar, les probabilitats femenines disminueixen un 31% (passa de 0,00143 a 0,00099) i les masculines augmenten un 48% (passa de 0,00393 a 0,00581). En el cas de “futbolista”, les probabilitats femenines disminueixen un 42% (passa de 0,00143 a 0,00082) i les masculines augmenten un 29% (passa de 0,00393 a 0,00508).

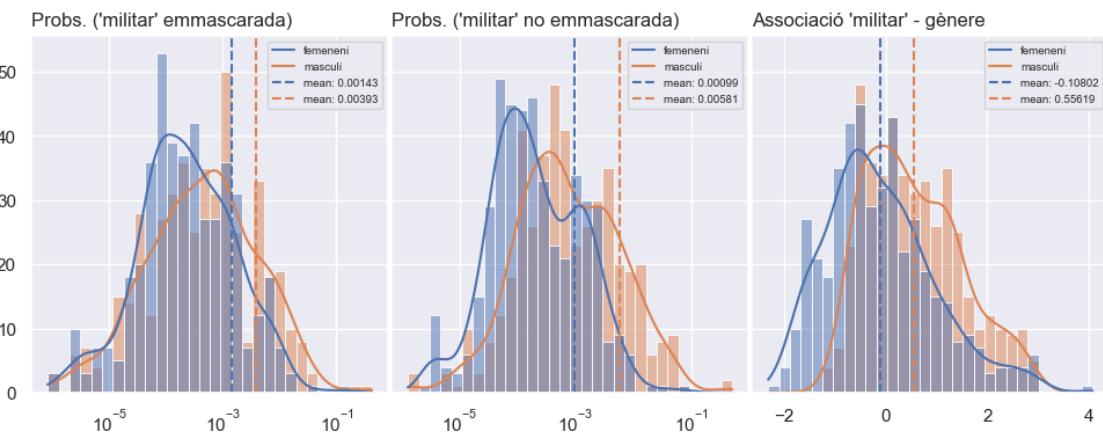


Figura 25 - Probabilitats i associació gènere-professió (militar).

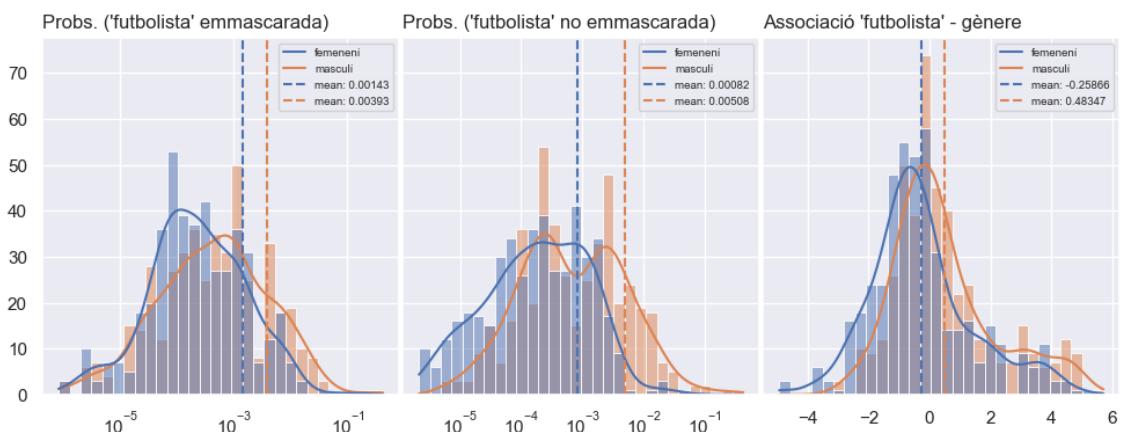


Figura 26 - Probabilitats i associació gènere-professió (futbolista).

La **taula 34**, mostra les diferències de representació trobades entre els gèneres masculí i femení, en diferents fases del treball. Es pot comprovar que en tots els casos, la ratio sempre és superior en favor del gènere masculí.

Procés	Femení	Masculí	Masc/Fem
1 Mitjana percentatges homes/dones professions Wikidata	24%	74%	3,17
2 Professions reconegudes BERTa (tokenització)	1	38	38

3	Professions amb marca de gènere	8	50	6,25
4	Probabilitats mitjanes de professions amb marca de gènere (gènere emmascarat)	0,00027	0,00100	3,7
5	Probabilitats mitjanes de paraules relatives al gènere (professions emmascarades)	0,00143	0,00393	2,75
5	Probabilitats mitjanes de paraules relatives al gènere (professions no emmascarades)	0,00193	0,00544	2,81
6	Probabilitats de paraules relatives al gènere per professions (no emmascarades)	1	48	48
7	Associació gènere – professió (sense marca de gènere)	4	45	11,25

Taula 34 - Resum de les principals diferències de representació entre gèneres.

5. Conclusions i treballs futurs

5.1 Conclusions principals

- El mètode Bartl no es pot replicar en model BERTa.

La gran diferència entre el llenguatge català i l'anglès, és la presència del gènere gramatical que fa que el **mètode seleccionat no sigui apte**. La presència del gènere gramatical afecta la concordança de l'oració té molta més influència en les associacions de paraules que no pas els possibles biaixos que puguin existir. Per aquest motiu s'ha hagut d'adaptar el mètode per les professions amb marca de gènere gramatical.

Les marques de gènere gramatical del llenguatge (professions amb marca de gènere – jutge/jutgessa) i la **concordança gramatical** amb la resta de paraules de l'oració tenen una influència major que l'existència o no de biaixos de gènere.

Les probabilitats de les professions amb gènere gramatical femení (jutgessa) en frases amb el gènere emmascatat són molt més baixes que les professions masculines (3,7 vegades superior, **Figura 7**). Per tant, si en frases on el gènere no està emmascatat aquestes probabilitats s'igualen, fa que l'associació gènere-professió sigui major, ja que augmenten en una proporció més gran.

- En el vocabulari de BERTa, el nombre de paraules en gènere grammatical masculí és superior al femení.

Hi ha menys paraules amb gènere gramatical femení que masculí. S'ha pogut comprovar amb la selecció de professions que de les **102 professions descartades que no els hi correspon un token** pel model BERTa, **37 es corresponen amb paraules femenines** i **1 paraula és masculina** (en 64 registres no es reconeix ni el terme masculí ni el femení). Aquesta diferència posa en relleu que, en les professions en qüestió, les de gènere femení són tan poc freqüents que no tenen un toquen i, per tant, no s'han pogut analitzar.

5.2 Conclusions secundàries

- A través de les associacions entre gènere i professió testejades al model BERTa, s'ha comprovat com existeix també un biaix d'estereotip.
- S'ha detectat un biaix d'infrarepresentació del gènere femení, amb les dades Wikidata, que probablement es reflecteix en l'observat en el model BERTa.
- A l'hora de determinar si els models són aptes o no, no és suficient amb les mètriques que comparen els resultats amb les dades utilitzades per entrenar el model perquè els biaixos de les mateixes també es reflecteixin als models.

- S'han de definir noves mètriques que incloguin la mesura de biaixos per determinar si els models són aptes o no.

5.3 Objectius assolits

Tot i que no s'ha definit una mètrica específica de mesura de biaixos de gènere, s'ha assolit l'objectiu principal de l'estudi de l'existència i quantificació de biaixos de gènere, remarcant les principals diferències del comportament del model per cadascun dels gèneres.

Pel que fa als objectius específics, no s'ha aconseguit l'objectiu específic de la **proposta de mesures per mitigar aquests biaixos**, més enllà de remarcar la necessitat de definir nous mètodes de mesures de rendiment dels models, a banda de les mètriques clàssiques que comparen els models amb les dades. A més, ha mancat trobar explorar millor els estudis sobre biaixos de gènere en models de llenguatges amb marques de gènere gramatical, i estudiar altres mesures d'associació que la proposada per Kurita et al. (2019) [2] i l'adaptació posterior de Bartl et al. (2020)[1].

5.4 Competència de “Compromís ètic i global”.

Pel que fa a la competència de “Compromís ètic i global”, aquest treball s'ha alineat principalment amb la dimensió de “Diversitat i drets humans” i, concretament, s'han adoptat els objectius de desenvolupament sostenible per al 2030 de l'ONU d' “Igualtat de gènere” (ODS 5) i de “Reducció de desigualtats” (ODS 10). També ha tingut un impacte en la dimensió de “Comportament ètic i de responsabilitat social”, alertant de les conseqüències de la utilització d'una intel·ligència artificial poc ètica. Així i tot, pel que fa a la **perspectiva de gènere**, en aquest estudi es fa un **tractament binari**, conscients que no inclou persones que s'identifiquen com no binàries que pot crear perjudicis de representació.

5.5 Línies de futur

Aquest treball ha obert moltes línies de futur, començant per l'objectiu específic no assolit de “**Proposar mesures per mitigar els biaixos**” incident en les dades i en les mètriques de rendiment de models. També, tenint en compte els **biaixos d'infrarepresentació** detectats, fer una anàlisi de les paraules amb gènere gramatical i comprovar si realment existeixen més paraules amb gènere masculí que femení, com ha semblat indicar amb la tokenització de les professions.

Finalment, obrir noves vies d'investigació amb l'ampliació de diferents tipus de plantilles per ampliar la varietat de contextos, incorporar noves professions, o paraules susceptibles de contenir biaixos com podrien ser per exemple trets de personalitat o emocions.

Traductor Automàtic

Text d'entrada

A father and his son are in a car accident. The father dies at the scene and the son is rushed to the hospital. At the hospital the surgeon looks at the boy and says "I can't operate on this boy, he is my son." How can this be?

Traducció

Un pare i el seu fill estan en un accident de cotxe. El pare mor en l'escena i el fill és traslladat de pressa a l'hospital. A l'hospital el cirurgià mira el noi i diu: "No puc operar amb aquest noi, és el meu fill". Com pot ser?

IA ètica i lliure de biaixos

cirurgiana

6. Glossari

AINA – Projecte impulsat pel Departament de la Vicepresidència i de Polítiques Digitals i Territori del Govern de Catalunya, que té com a objectiu generar recursos i models de la llengua catalana per a ser utilitzats en aplicacions basades en intel·ligència artificial i processament del llenguatge natural PLN.

BERT - Bidirectional Encoder Representations from Transformers (model del llenguatge basat en deep learning)

BERTa - Adaptació al català del model RoBERTa

BSC - Barcelona Supercomputing Center

ELMo - Embeddings from Language Model (model del llenguatge basat en deep learning).

GPT - Generative Pre-trained Transformer (model del llenguatge basat en deep learning).

IA – Intel·ligència Artificial.

MLM - Masked Language Modeling (modelatge del llenguatge emmascarat).

PLN - Processament del llenguatge natural PLN (Natural Language Processing - NLP)

RNN - Recursive Neural Network (Xarxa neuronal recurrent).

RoBERTa - Robustly Optimized BERT Pretraining Approach (model del llenguatge basat en deep learning).

Semàntica - estudia el significat de les paraules.

Sintaxi - part de la gramàtica que estudia les regles de combinació de paraules.

SPARQL - SPARQL Protocol and RDF Query Language (acrònim recursiu).

Transfer learning - Aprofitar de models pre-entrenats per unes tasques per ser reutilitzats en altres tasques.

Wikidata - base de dades lliure de dades estructurades per projectes com ara Viquipèdia.

Word2vect - algorisme per la creació d'un model del llenguatge basat en la predicció de paraules.

Word embedding - Representació de vectors de paraules

7. Bibliografia

- [1] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K. W. Chang, "Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints," *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, pp. 2979–2989, 2017, doi: 10.18653/V1/D17-1323.
- [2] M. R. Costa-jussà, "An analysis of gender bias studies in natural language processing," *Nature Machine Intelligence* 2019 1:11, vol. 1, no. 11, pp. 495–496, Oct. 2019, doi: 10.1038/s42256-019-0105-5.
- [3] M. Bartl, M. Nissim, and A. Gatt, "Unmasking Contextual Stereotypes: Measuring and Mitigating BERT's Gender Bias," Oct. 2020, doi: 10.48550/arxiv.2010.14534.
- [4] J. Armengol-Estabé *et al.*, "Are Multilingual Models the Best Choice for Moderately Under-resourced Languages? A Comprehensive Assessment for Catalan," *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 4933–4946, 2021, doi: 10.18653/V1/2021.FINDINGS-ACL.437.
- [5] Z. Sadeghi, J. L. McClelland, and P. Hoffman, "You shall know an object by the company it keeps: An investigation of semantic representations derived from object co-occurrence in visual scenes," *Neuropsychologia*, vol. 76, pp. 52–61, Sep. 2015, doi: 10.1016/J.NEUROPSYCHOLOGIA.2014.08.031.
- [6] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global Vectors for Word Representation," *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pp. 1532–1543, 2014, doi: 10.3115/V1/D14-1162.
- [7] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, Jan. 2013, doi: 10.48550/arxiv.1301.3781.
- [8] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with Subword Information," *Trans Assoc Comput Linguist*, vol. 5, pp. 135–146, Jul. 2016, doi: 10.48550/arxiv.1607.04606.
- [9] T. Mikolov, W. Yih, and G. Zweig, "Linguistic Regularities in Continuous Space Word Representations." Association for Computational Linguistics, pp. 746–751, 2013. Accessed: Jan. 13, 2023. [Online]. Available: <https://aclanthology.org/N13-1090>
- [10] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving Language Understanding by Generative Pre-Training," 2018.

- [11] A. Vaswani *et al.*, “Attention Is All You Need,” *Adv Neural Inf Process Syst*, vol. 2017-December, pp. 5999–6009, Jun. 2017, doi: 10.48550/arxiv.1706.03762.
- [12] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, vol. 1, pp. 4171–4186, Oct. 2018, doi: 10.48550/arxiv.1810.04805.
- [13] Y. Liu *et al.*, “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” Jul. 2019, doi: 10.48550/arxiv.1907.11692.
- [14] A. Conneau *et al.*, “Unsupervised Cross-lingual Representation Learning at Scale,” Nov. 2019, doi: 10.48550/arxiv.1911.02116.
- [15] L. Martin *et al.*, “CamemBERT: a Tasty French Language Model,” pp. 7203–7219, Jul. 2020, doi: 10.18653/V1/2020.ACL-MAIN.645.
- [16] W. de Vries, A. van Cranenburgh, A. Bisazza, T. Caselli, G. van Noord, and M. Nissim, “BERTje: A Dutch BERT Model,” Dec. 2019, doi: 10.48550/arxiv.1912.09582.
- [17] A. Virtanen *et al.*, “Multilingual is not enough: BERT for Finnish,” Dec. 2019, doi: 10.48550/arxiv.1912.07076.
- [18] A. Mukerjee, R. Biswas, K. Deb, and A. P. Mathur, “Multi-objective Evolutionary Algorithms for the Risk–return Trade–off in Bank Loan Management,” *International Transactions in Operational Research*, vol. 9, no. 5, pp. 583–597, Sep. 2002, doi: 10.1111/1475-3995.00375.
- [19] L. Cohen, Z. C. Lipton, and Y. Mansour, “Efficient candidate screening under multiple tests and implications for fairness,” *Leibniz International Proceedings in Informatics, LIPIcs*, vol. 156, May 2019, doi: 10.48550/arxiv.1905.11361.
- [20] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, “Machine Bias: there’s software used across the country to predict future criminals. And it’s biased against blacks.,” *ProPublica*, 2016. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> (accessed May 14, 2022).
- [21] M. Kaneko, A. Imankulova, D. Bollegala, and N. Okazaki, “Gender Bias in Masked Language Models for Multiple Languages,” *NAACL 2022 - 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, pp. 2740–2750, 2022, doi: 10.18653/V1/2022.NAACL-MAIN.197.
- [22] T. Bolukbasi, K. W. Chang, J. Zou, V. Saligrama, and A. Kalai, “Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings,” *Adv Neural Inf Process Syst*, pp. 4356–4364, Jul. 2016, doi: 10.48550/arxiv.1607.06520.

- [23] T. Sun *et al.*, “Mitigating Gender Bias in Natural Language Processing: Literature Review,” *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pp. 1630–1640, 2019, doi: 10.18653/V1/P19-1159.
- [24] A. Caliskan, J. J. Bryson, and A. Narayanan, “Semantics derived automatically from language corpora contain human-like biases,” *Science* (1979), vol. 356, no. 6334, pp. 183–186, Aug. 2016, doi: 10.1126/science.aal4230.
- [25] M. Nissim, R. van Noord, and R. van der Goot, “Fair is better than sensational: Man is to doctor as woman is to doctor,” *Computational Linguistics*, vol. 46, no. 2, pp. 487–497, 2020, doi: 10.1162/COLI_a_00379.
- [26] S. Kiritchenko and S. M. Mohammad, “Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems,” *NAACL HLT 2018 - Lexical and Computational Semantics, SEM 2018, Proceedings of the 7th Conference*, pp. 43–53, May 2018, doi: 10.48550/arxiv.1805.04508.
- [27] J. Zhao, T. Wang, M. Yatskar, R. Cotterell, V. Ordonez, and K.-W. Chang, “Gender Bias in Contextualized Word Embeddings,” 2019.
- [28] C. May, A. Wang, S. Bordia, S. R. Bowman, and R. Rudinger, “On measuring social biases in sentence encoders,” in *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 2019, vol. 1, pp. 622–628. doi: 10.18653/v1/n19-1063.
- [29] K. Kurita, N. Vyas, A. Pareek, A. W. Black, and Y. Tsvetkov, “Measuring Bias in Contextualized Word Representations,” 2019. [Online]. Available: <https://bit.ly/2EkJwh1>
- [30] Y. Wu *et al.*, “Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation,” Sep. 2016, doi: 10.48550/arxiv.1609.08144.
- [31] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language Models are Unsupervised Multitask Learners,” 2018.
- [32] M. Nadeem, A. Bethke, and S. Reddy, “StereoSet: Measuring stereotypical bias in pretrained language models,” *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pp. 5356–5371, Apr. 2020, doi: 10.48550/arxiv.2004.09456.
- [33] S. Kiritchenko and S. M. Mohammad, “Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems,” *NAACL HLT 2018 - Lexical and Computational Semantics, SEM 2018, Proceedings of the 7th Conference*, pp. 43–53, May 2018, doi: 10.48550/arxiv.1805.04508.
- [34] A. Caliskan, J. J. Bryson, and A. Narayanan, “Semantics derived automatically from language corpora contain human-like biases,” *Science* (1979), vol. 356, no. 6334, pp. 183–186, Apr. 2017, doi: 10.1126/SCIENCE.AAL4230/SUPPL_FILE/CALISKAN-SM.PDF.

- [35] T. Davidson, D. Bhattacharya, and I. Weber, “Racial Bias in Hate Speech and Abusive Language Detection Datasets,” pp. 25–35, May 2019, doi: 10.48550/arxiv.1905.12516.
- [36] M. Sap, D. Card, S. Gabriel, Y. Choi, and N. A. Smith, “The Risk of Racial Bias in Hate Speech Detection,” *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pp. 1668–1678, 2019, doi: 10.18653/V1/P19-1163.

8. Annexos

Annex A – Professions no reconegudes amb un sol token.

Annex B – Professions utilitzades en l'estudi.

Annex A – Professions no reconegudes amb un sol token.

La taula 18 mostra les professions que o bé la paraula femenina (37 professions) o bé la paraula masculina (1 professió), o bé cap de les dues (64 professions), no es troben al diccionari del model BERTa i, per tant, no es codifica en un sol token.

Es pot consultar l'entrada de la professió a *Wikidata* a l'URL <https://www.wikidata.org/wiki/> i la seva ID. Per exemple, la pàgina d'*agricultor/agricultora*, amb l'identificador Q131512, es pot veure a <https://www.wikidata.org/wiki/Q131512>.

No es reconeix ni el terme masculí ni el femení.

#	Wikidata ID	Mascúli	Femení
1	Q131512	agricultor	agricultora
2	Q16308156	anglicista	anglicista
3	Q266569	animador	animadora
4	Q4773904	antropòleg	antropòloga
5	Q3621491	arqueòleg	arqueòloga
6	Q1382355	arquer	arquera
7	Q1643514	arranjador	arranjadora
8	Q635734	arxiver	arxivera
9	Q11631	astronauta	astronauta
10	Q11063	astrònom	astrònoma
11	Q864380	biògraf	biògrafa
12	Q2919046	bioquímic	bioquímica
13	Q8246794	bloguer	bloguera
14	Q11338576	boxejador	boxejadora
15	Q7541856	ceramista	ceramista
16	Q130857	discjòquei	discjòquei
17	Q15839134	ecòleg	ecòloga
18	Q165029	electricista	electricista
19	Q3055126	entomòleg	entomòloga
20	Q9149093	escalador	escaladora
21	Q2707485	escenògraf	escenògrafa
22	Q4144610	esquiador	esquiadora
23	Q372436	estadista	estadista
24	Q1371378	etnòleg	etnòloga
25	Q3126128	genetista	genetista
26	Q520549	geòleg	geòloga
27	Q2599593	germanista	germanista
28	Q13381572	gimnasta	gimnasta
29	Q2640827	ginecòleg	ginecòloga
30	Q11569986	gravador	gravadora
31	Q644687	il·lustrador	il·lustradora
32	Q15924224	internista	internista
33	Q205375	inventor	inventora
34	Q2519376	joier	joiera
35	Q846750	joquei	joquei
36	Q6665249	judoka	judoka
37	Q9017214	karateka	karateka
38	Q822146	lletrista	lletrista
39	Q935666	maquillador	maquilladora
40	Q3332711	medievalista	medievalista
41	Q3779582	microbiòleg	microbiòloga
42	Q18545066	monologuista	monologuista
43	Q14915627	musicòleg	musicòloga
44	Q10843402	nedador	nedadora
45	Q6625963	novel·lista	novel·lista
46	Q1225716	ornitòleg	ornitòloga
47	Q1662561	paleontòleg	paleontòloga
48	Q13219587	patinador	patinadora
49	Q1919436	pediatre	pediatra
50	Q13382566	piragüista	piragüista
51	Q1900167	psicoterapeuta	psicoterapeuta
52	Q211346	psiquiatre	psiquiatra
53	Q476246	regatista	regatista
54	Q13382576	remer	remera
55	Q2504617	romanista	romanista
56	Q13382603	saltador	saltadora
57	Q12800682	saxofonista	saxofonista
58	Q10379007	teatròleg	teatròloga
59	Q13381863	tirador	tiradora
60	Q23892384	trampolinista	trampolinista
61	Q15306067	triatleta	triatleta
62	Q4009406	velocista	velocista
63	Q17125263	youtuber	youtuber
64	Q350979	zoòleg	zoòloga

No es reconeix el terme masculí.

#	Wikidata ID	Mascúli	Femení	Token ID (fem)
1	Q186360	infermer	infermera	26849

No es reconeix el terme femení.

#	Wikidata ID	Mascúli	Femení	Token ID (masc)
1	Q21281706	administrador	administradora	29509
2	Q42973	arquitecte	arquitecta	18685
3	Q7019111	assistent	assistenta	18802
4	Q806798	banquer	banquera	43648
5	Q864503	biòleg	biòloga	43807
6	Q107711	bomber	bombera	40199
7	Q1735282	caixer	caixera	35481
8	Q488205	cantautor	cantautora	22098
9	Q774306	cirurgià	cirurgiana	36889
10	Q2490358	coreògraf	coreògrafa	49845
11	Q4439155	corredor	corredora	12884
12	Q214917	dramaturg	dramaturga	25305
13	Q43845	empresari	empresària	4583
14	Q81096	enginyer	enginyera	19110
15	Q628099	entrenador	entrenadora	12037
16	Q1281618	escultor	escultora	34654
17	Q13418253	filòleg	filòloga	31196
18	Q4964182	filòsof	filòsofa	12643
19	Q154549	fuster	fusteria	45636
20	Q901402	geògraf	geògrafa	48411
21	Q27214348	inspector	inspectora	38498
22	Q16358610	instructor	instructora	27287
23	Q998550	llibreter	llibretera	46059
24	Q12369333	lluitador	lluitadora	35770
25	Q2722764	locutor	locutora	46540
26	Q755070	narrador	narradora	27398
27	Q838811	pàges	pàgesa	15786
28	Q152002	pastor	pastora	20190
29	Q1231865	pedagog	pedagoga	36339
30	Q1238570	politòleg	politòloga	44488
31	Q5482740	programador	programadora	47655
32	Q2252262	raper	radera	38119
33	Q3455803	realitzador	realitzadora	34318
34	Q42909	reporter	reportera	41494
35	Q250867	sacerdot	sacerdotessa	18926
36	Q2306091	sociòleg	sociòloga	31592
37	Q1234713	tèòleg	tèologa	46537
38	Q216541	xofer	xofera	42093

Taula 35 - Annex A - Professions no reconegudes pel model BERTa.

Annex B – Professions utilitzades en l'estudi

A continuació es mostren les 107 professions utilitzades en l'estudi un cop netejades, amb els valors extrets de Wikidata de total de dones i d'homes i el percentatge del total de dones respecte del total.

#	Wikidata ID	Femení	Masculí	Dones	Homes	Total	% dones
1	Q33999	actriu	actor	153.097	198.409	351.506	44%
2	Q82955	política	polític	64.347	286.607	350.954	18%
3	Q3665646	esportista	esportista	52.709	273.905	326.614	16%
4	Q937857	futbolista	futbolista	45.536	259.998	305.534	15%
5	Q36180	escriptora	escriptor	68.101	147.372	215.473	32%
6	Q1622272	professora	professor	38.667	135.719	174.386	22%
7	Q1930187	periodista	periodista	29.428	88.684	118.112	25%
8	Q483501	artista	artista	41.141	69.473	110.614	37%
9	Q2526255	directora	director	15.626	83.438	99.064	16%
10	Q177220	cantant	cantant	43.759	51.667	95.426	46%
11	Q1028181	pintora	pintor	22.834	66.189	89.023	26%
12	Q201788	historiadora	historiador	21.226	67.492	88.718	24%
13	Q36834	compositora	compositor	11.939	68.994	80.933	15%
14	Q639669	música	músic	8.904	53.149	62.053	14%
15	Q1650915	investigadora	investigador	16.725	38.384	55.109	30%
16	Q49757	poeta	poeta	14.416	40.063	54.479	26%
17	Q3282637	productora	productor	9.061	42.613	51.674	18%
18	Q28389	guionista	guionista	9.500	39.807	49.307	19%
19	Q40348	advocada	advocat	7.161	35.609	42.770	17%
20	Q11513337	atleta	atleta	15.146	24.777	39.923	38%
21	Q482980	autora	autor	10.756	28.936	39.692	27%
22	Q33231	fotògrafa	fotògraf	8.782	29.176	37.958	23%
23	Q2309784	ciclista	ciclista	7.158	27.013	34.171	21%
24	Q39631	metgessa	metge	6.039	27.284	33.323	18%
25	Q333634	traductora	traductor	11.572	20.976	32.548	36%
26	Q47064	militar	militar	761	31.332	32.093	2%
27	Q901	científica	científic	5.741	26.168	31.909	18%
28	Q876864	editora	editor	7.454	23.823	31.277	24%
29	Q188094	economista	economista	4.244	26.824	31.068	14%
30	Q121594	catedràtica	catedràtic	5.835	21.924	27.759	21%
31	Q4610556	model	model	22.316	4.041	26.357	85%
32	Q10349745	pilot	pilot	1.779	23.645	25.424	7%
33	Q947873	presentadora	presentador	10.645	13.920	24.565	43%
34	Q193391	diplomàtica	diplomàtic	2.517	21.923	24.440	10%
35	Q5322166	dissenyadora	dissenyador	6.802	15.351	22.153	31%
36	Q185351	jurista	jurista	3.616	17.587	21.203	17%
37	Q212238	funcionària	funcionari	4.471	16.483	20.954	21%
38	Q169470	física	físic	1.658	18.785	20.443	8%
39	Q16533	jutgessa	jutge	3.264	16.913	20.177	16%
40	Q170790	matemàtica	matemàtic	2.321	16.494	18.815	12%
41	Q189290	oficial	oficial	581	17.956	18.537	3%
42	Q4263842	crítica	crític	3.458	14.359	17.817	19%
43	Q486748	pianista	pianista	4.325	12.430	16.755	26%
44	Q855091	guitarrista	guitarrista	1.211	14.837	16.048	8%
45	Q593644	química	químic	2.224	13.220	15.444	14%
46	Q131524	emprenedora	emprendedor	2.100	13.274	15.374	14%
47	Q15253558	activista	activista	7.614	7.210	14.824	51%
48	Q212980	psicòloga	psicòleg	4.714	8.876	13.590	35%
49	Q2374149	botànica	botànic	3.522	10.015	13.537	26%
50	Q3400985	acadèmica	acadèmic	3.971	9.154	13.125	30%
51	Q5716684	ballarina	ballarí	7.682	5.216	12.898	60%
52	Q14467526	lingüista	lingüista	3.631	8.775	12.406	29%
53	Q15296811	dibuixant	dibuixant	2.547	9.522	12.069	21%

54	Q10833314	tenista	tenista	5.180	6.030	11.210	46%
55	Q82594	informàtica	informàtic	1.155	8.232	9.387	12%
56	Q182436	bibliotecària	bibliotecari	4.240	4.592	8.832	48%
57	Q859528	àrbitre	àrbitre	594	6.240	6.834	9%
58	Q2462658	mànager	mànager	1.312	5.319	6.631	20%
59	Q15627169	sindicalista	sindicalista	1.106	5.454	6.560	17%
60	Q245068	còmic	còmic	1.376	5.141	6.517	21%
61	Q1259917	violinista	violinista	1.419	3.633	5.052	28%
62	Q4991371	soldat	soldat	123	4.607	4.730	3%
63	Q386854	bateria	bateria	154	4.272	4.426	3%
64	Q713200	intèrpret	intèrpret	1.618	2.314	3.932	41%
65	Q780596	comissària	comissari	1.604	2.327	3.931	41%
66	Q384593	policia	policia	307	3.429	3.736	8%
67	Q2135469	assessora	assessor	763	2.736	3.499	22%
68	Q2251335	mestra	mestre	1.052	2.175	3.227	33%
69	Q974144	educadora	educador	1.210	2.016	3.226	38%
70	Q765778	organista	organista	425	2.666	3.091	14%
71	Q17276321	diputada	diputat	444	2.498	2.942	15%
72	Q1414443	cineasta	cineasta	790	2.094	2.884	27%
73	Q462390	docent	docent	875	1.984	2.859	31%
74	Q215536	comerciant	comerciant	209	2.634	2.843	7%
75	Q202883	veterinària	veterinari	457	2.152	2.609	18%
76	Q2986228	comentarista	comentarista	406	2.196	2.602	16%
77	Q15978655	consultora	consultor	566	1.773	2.339	24%
78	Q105186	farmacèutica	farmacèutic	623	1.692	2.315	27%
79	Q2732142	estadística	estadístic	443	1.737	2.180	20%
80	Q1086863	columnista	columnista	703	1.464	2.167	32%
81	Q1423891	ministra	ministre	192	1.907	2.099	9%
82	Q27349	dentista	dentista	343	1.463	1.806	19%
83	Q519076	agent	agent	386	1.343	1.729	22%
84	Q9379869	conferenciant	conferenciant	1.280	421	1.701	75%
85	Q708492	regidora	regidor	241	1.458	1.699	14%
86	Q3499072	xef	xef	439	1.141	1.580	28%
87	Q156839	cuinera	cuiner	376	1.071	1.447	26%
88	Q30185	alcaldessa	alcalde	117	1.077	1.194	10%
89	Q326653	comptable	comptable	218	940	1.158	19%
90	Q140686	presidenta	president	235	892	1.127	21%
91	Q80687	secretària	secretari	475	528	1.003	47%
92	Q1155838	corresponsal	corresponsal	201	727	928	22%
93	Q12713481	obrera	obrer	99	783	882	11%
94	Q485178	analista	analista	171	561	732	23%
95	Q832136	gerent	gerent	126	559	685	18%
96	Q1039099	guia	guia	161	513	674	24%
97	Q327029	mecànica	mecànic	19	483	502	4%
98	Q16532929	administrativa	administratiu	113	203	316	36%
99	Q28789517	guarda	guarda	45	267	312	14%
100	Q327321	paleta	paleta	3	191	194	2%
101	Q24716636	voluntària	voluntari	79	111	190	42%
102	Q2986261	comercial	comercial	19	141	160	12%
103	Q157195	cambrera	cambrer	52	60	112	46%
104	Q2180295	cartera	carter	16	92	108	15%
105	Q2961580	taxista	taxista	8	89	97	8%
106	Q1058617	detectiu	detectiu	6	88	94	6%
107	Q25141651	doctora	doctor	20	5	25	80%