eman ta zabal zazu

Universidad del País Vasco

Euskal Herriko Unibertsitatea

Study of a Metric for Measuring Gender Bias in BERT Language Models

Author: Elvira González Hernández

Advisors: Olatz Arbelaitz, Olatz Pérez de Viñaspre

hap/lap

Hizkuntzaren Azterketa eta Prozesamendua
Language Analysis and Processing

Final Thesis

September 2022

Departments: Computer Systems and Languages, Computational Architectures and Technologies, Computational Science and Artificial Intelligence, Basque Language and Communication, Communications Engineer.

A word after a word after a word is power.

\- Margaret Atwood

Resumen

Desde su creación, los modelos del lenguaje BERT están siendo implementados en multitud de plataformas que dan servicio a millones de usuarios. Debido a su creciente popularidad, se empezó a dar importancia al hecho de crear sistemas éticos dentro del campo del procesamiento del lenguaje natural, sobre todo teniendo en cuenta los perjuicios que un sistema que no sea justo e imparcial puede producir en algunos grupos de la sociedad. Por este motivo, se está investigando cada vez más sobre técnicas para la detección y reducción del sesgo de género. En este trabajo se va a analizar una métrica para la medición del sesgo de género estudiando la asociación entre referentes con marca de género y profesiones. Se cuestionarán los resultados obtenidos con la misma métrica en trabajos anteriores y, llevando a cabo un análisis más exhaustivo se examinarán las limitaciones que tiene la métrica.

Los experimentos se llevarán a cabo en tres idiomas, inglés, euskera y español; en los tres modelos BERT monolingües correspondientes a cada uno de ellos: BERT base, BERTeus y BETO. La variedad lingüística de los tres idiomas en cuanto al género gramatical, junto con el análisis más detallado de la métrica ayudará a obtener interesantes conclusiones sobre las limitaciones de la métrica.

Palabras clave: sesgo de género, BERT, métrica, modelos del lenguaje

Abstract

Since the creation of language models such as BERT, they are being deployed widely as services on platforms to serve millions of users. With their increasing popularity, the fairness of NLP systems and algorithms is a subject of great interest nowadays given the harms an unethical system can cause. That is why researchers have been interested in the development of techniques for detection and mitigation of bias. In this work, a previously proposed metric for measuring gender bias by studying associations between gender-denoting referents and names of professions will be analysed. The accuracy of previous results will be questioned, and a deeper analysis of the metric will demonstrate that the metric has some flaws and limitations in the way it represents gender bias.

The experiments will be carried out for three languages, English, Basque and Spanish, and its corresponding monolingual BERT models: BERT base, BERTeus and BETO. The fact that the three languages are very different linguistically, especially regarding grammatical gender, together with a thorough analysis of the metric will reveal some interesting conclusions about the metric's limitations.

Keywords: gender bias, BERT, metric, language models

Acknowledgements

I would like to express my warmest gratitude to my supervisors, Olatz *eta* Olatz, for their invaluable support and guidance during this project. Thanks for always cheering me up, for being so understanding and patient and for helping me to give shape to what at times seemed like just a huge bunch of numbers. I would also like to express my gratitude to Elena Martínez Caro, linguistics professor at Universidad Complutense de Madrid, thanks for helping me get into this master's program, for sharing your passion for linguistics and inspiring your students to grow.

Many thanks to my friends, who are a constant support in my life and are always there for me, regardless of being physically close or not. I am also grateful to all the very nice people I met during the master, who made 2021 one of the best years ever. To my family, for always being proud of my achievements. Especially to my grandparents, who will always have a privileged place in my heart; you made me the person I am today and I will be always grateful to you both.

Finally, special thanks to my favourite person, who also appeared in my life during the master. Thanks for just being there unconditionally and for always making me laugh so much. This document would not be half as elegant and nice without your intervention and feedback. Thanks for patiently listening to my monologues about this thesis.

Also, shout-out to the cat, for always keeping me company no matter what :)

Haplap has changed me so much and has brought to my life so many good things: incredible people, knowledge, a new job, a new city to live in... I am glad to be able to give closure to this cycle with a project I am very proud of.

Contents

List of Figures

List of Tables

1 Introduction

The field of Natural Language Processing (NLP) experimented a turning point with the creation of BERT language models. They became a state of art in various NLP tasks such as question answering, Named Entity Recognition (NER), Natural Language Inference (NLI), text classification, etc. Language models are learned from massive text corpora using variants of language modeling objective, that is, correctly predicting a word given its surrounding context (Nadeem et al., 2020). They are being deployed widely as services on platforms to serve millions of users. With their increasing popularity, concerns arise about the fairness of these models and how ethically they are being built. Since they learn from extensive text corpora, there is a danger that stereotypical biases in the real world are reflected in these models. As a consequence, a large body of works analyzing bias in NLP systems has appeared in recent years, including works on bias in embedding spaces (Kumar et al., 2020), and for many tasks such as coreference resolution (Cao and Daumé, 2021), machine translation (Costa-jussà et al., 2022), sentiment analysis (Asyrofi et al., 2021) or hate speech detection (Park et al., 2018).

This project intends to contribute to the literature about gender bias and its presence in BERT models. Gender bias is the preference of one gender over the others, based on prejudices (Moss-Racusin et al., 2012). They cause harms and have real world consequences in individuals and society as a whole, for instance, automatic resume filtering systems giving preference to male applicants when the only distinguishing factor is the applicants' gender (Sun et al., 2019) or biased text generation as addressed in (Tatman, 2017) in YouTube's automatically-generated captions; just to mention a couple of examples. So this is one of the main motivations of this project, to contribute to the research regarding the fairness of NLP systems to help raise awareness about the potential discrimination they may reproduced and, hopefully, mitigate it.

More specifically, the main goal of the project is to analyse a metric for measuring gender bias proposed by (Bartl et al., 2020), and later revisited by (Azpillaga Rivera, 2021) focusing on three languages and their corresponding pre-trained monolingual BERT models: BERT base (English), BERTeus (Basque) and BETO (Spanish). The metric measures gender bias by studying associations between gender-denoting target words and names of professions, comparing the findings with real-world workforce statistics. It was concluded to successfully quantify gender bias in BERT for English in (Bartl et al., 2020), demonstrating that there is indeed bias ingrained in the model; but it did not work the same neither for Basque or Spanish (Azpillaga Rivera, 2021). So, the objective is to dig deeper in the logic of the metric and analyze the elements that could condition the results it provides to find out why is it that the results do not show bias in the models for Basque or Spanish. For doing that, English will be revisited first, to see if we can extract more details about its results carrying out a more profound analysis. Then, the same will be done for Basque and Spanish, but carrying out some changes for the replication of the experiments, such as modifying the list of professions used.

So, assuming that there exists potential gender bias in BERT language models, these are the proposed hypotheses:

- For **Basque**, a neutral bias is expected, given that it is a language with few grammatical gender marks.
- For **Spanish**, bias should be present attending to the fact that it is a gender-marking language.

It will be seen that for Basque, even if some traces of neutrality are observed, the results show bias with a tendency towards the masculine gender. Regarding Spanish, even though it is strongly marked grammatically, the results showed no evidence of bias. For English, considering that the results in previous works showed the expected bias, after the analysis it was seen that this was not the case for all professions and that the metric has some limitations that may lead to question the results previously obtained. The overall results of this project also demonstrate the necessity of carrying out experiments with languages other than English –and especially minority languages such as Basque– because in most of the cases, conclusions cannot be extrapolated and it could lead to misleading statements and interpretations.

This document is organised as follows: Section 2 presents gender-based theory related to how gender is codified in the languages analysed, the definition of gender bias and the potential harms that developing unfair systems in the context of NLP can produce. Besides, the current techniques for measuring and mitigation of gender bias in NLP are reviewed. Then, in Section 3, the steps followed to carry out the experiments are addressed. After that, Section 4 contains the findings, analysis and discussion of results. It also contains important resources used in the experiments, such as the corpora and the list of professions. Lastly, Section 5 sums up the project and proposes some lines of future work.

Bias Statement

In this project, we study stereotypical associations between male and female gender and professional occupations in BERT language models. If a system systematically and by default associates certain professions with a specific gender, this creates a representational harm (Crawford, 2017) by perpetuating inappropriate stereotypes about what activities men and women are able, allowed or expected to perform. We focus on gender bias specifically, defined as the systematic unequal treatment on the basis of gender (Sun et al., 2019). While I am treating gender as binary in this study, I am aware that this does not include people who identify as non-binary, which can create representational harm (Blodgett et al., 2020).

2 Literature Review

2.1 Deep Learning in Natural Language Processing and the Development of Language Models

The onset of neural networks occurred in the 1940s, driven by the desire of using computers to do intelligent tasks. At first, the attempts were not very successful: computers weren't powerful enough, and the amount of digital data required was still scarce. Improvements made in these two areas over the next few decades, allowed significant performance gains of neural networks by the usage of deep learning techniques. Artificial intelligence, machine learning and deep learning are concepts that need to be differentiated. Artificial intelligence is the ability for a machine to perform rational and deductive processes of living beings, being machine learning a possible way of achieving this. Machine learning is, therefore, a branch of artificial intelligence that aims to develop computer programs capable of learning from experience, without being programmed to solve a particular task. Deep learning is a set of machine learning techniques based on computer models that mimic information processing in biological nervous systems, that is to say, artificial neural networks. Although in the early stages, deep learning was intended to simulate the human brain, today it refers to a more general principle in which knowledge of statistics and applied mathematics is used as a basis. In many fields, deep learning has managed to overcome other methods for tasks like computer vision, bio-informatics, pattern recognition, language processing, etc.

In natural language processing, language models work by assigning a probability to a sequence of words, that allows them to differentiate between similar words and phrases. This probability is usually used to predict what the next word can be, by using the information of the previous part of the sentence, although this is not always the case. Language models are very useful in different tasks of language processing, especially those based on text generation. By training language models based on neural networks (Neural language models) using massive amount of text data, the number of different words known –the vocabulary– increases. When the vocabulary increases, the number of different sentences that can occur increases exponentially. This can cause problems in large dimension spaces. To avoid this problem in neural networks, word representations –*embeddings*– are used.

2.1.1 Representations of language models: Embeddings

We call embeddings to vectors formed by real numbers that assign an abstract representation to words or phrases (Almeida and Xexéo, 2019). These vectors encode each word in a multi-dimensional vector space, as if they were coordinates of a map where each word takes a concrete position. This allows us to easily measure similarities and relations between words through mathematical operations. Embeddings can be used to solve many language processing tasks. Given their usefulness, there are several pre-processed embeddings available, such as fastText Caliskan et al. (2017b), which take into account all appearances of a word in certain texts –typically Wikipedia and Common Crawl[1]–, with-

[1] `https://commoncrawl.org/`

out looking at the different meanings that words can have: these are called *Static Word Embeddings*.

Static word embeddings do not differentiate between the different senses of a word so, for tasks that need to rely in the context, context-dependent embeddings (*Contextual Embeddings*) are used. In this type of embeddings, the different meanings of a word are taken into account when creating them; different contexts generate different representations for the same word. When talking about the context a word appears in, the idea is a piece of text of fixed length and, in the center, the word we are interested in. Although embeddings could be generated from a hand-labeled corpus, it is not an easy task to create *good* embeddings for all the possible senses of all existing words in a language, as for each sense a large number of examples are needed, and many unusual senses do not appear enough times.

2.1.2 Transformers

A Recurrent Neural Network (RNN) is a type of neural network that is able to process a sequence of arbitrary length by recursively applying a transition function to its internal hidden state vector h_t of the input sequence. The activation of the hidden state h_t at time-step t is computed as a function f of the current input symbol x_t and the previous hidden state h_{t-1} (Liu et al., 2016). This allows the RNN to have short-term memory to retain past information and, hence, uncover relationships between data points that are far from each other. However, this short-term memory is not enough when processing long sequences and, oftentimes, the output does not maintain information of the initial elements of the sentence. This can be a major problem, for example, when translating text between two different languages, as the first words of the original sentence are important for the quality of the translation. An example of this can be illustrated with languages where the subject is placed in first position, this information may be needed for the concordance with the verb and/or other sentence components. To solve this problem, a mechanism called *attention* is added, which allows the identification of the most important elements of the input sequence.

Transformers (Vaswani et al., 2017) are multi-layered deep learning architectures formed by stacking Transformer blocks on top of one another. Transformer blocks are characterized by a multi-head self-attention mechanism (attention to the other words in the same sequence) that are interleaved with nonlinear functions applied to individual vectors (Tay et al., 2020). They are designed to handle ordered sequences of data, such as a sentence written in natural language. Unlike RNNs, by using transformers it is no longer necessary to process sequences in an orderly manner. For example, when processing a text sentence, to process the end of the sentence it would not be necessary to have the beginning already processed. This allows transformers to be easily parallelized, reducing significantly the training cost. The architecture of a transformer can be seen in Figure 1. Although they were initially proposed for the task of machine translation, their usage has spread to many different tasks in recent years and has improved the state of art in most of them.

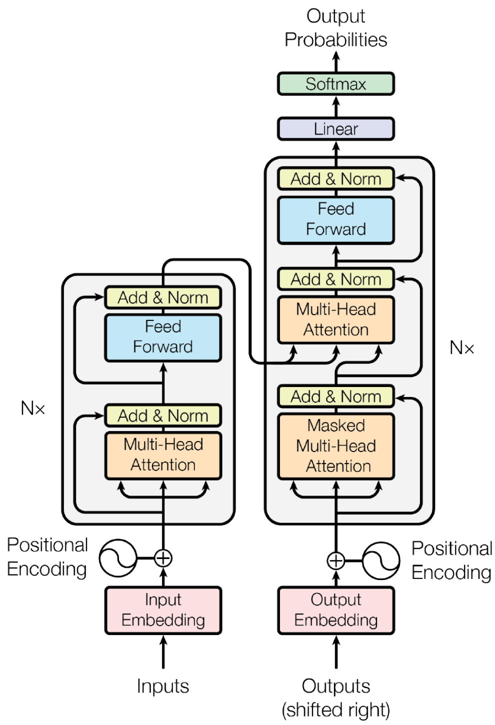

Figure 1: Architecture of a transformer. Source: Vaswani et al. (2017).

2.1.3 BERT Language Models

As said before, a language model is a statistical tool that analyzes the pattern of human language for the prediction of words or sequences of words. Before the creation of BERT (Devlin et al., 2018a), the existent language models were not context-dependent. BERT, however, takes into account the context in order to make predictions of words within a sentence. From the moment it was created, it has been growing rapidly and extended for other languages, as it was originally trained for English, including multilingual versions of it. It transformed the field of NLP, becoming "a ubiquitous baseline in NLP experiments in a little over a year" (Rogers et al., 2020).

As the goal of this project is to investigate if BERT language models learn and assimilate gender bias, as they are trained with human-made data, this section intends to give an overview of what BERT language models are, how do they work through the explanation of transformers architecture, and which are the ones that are going to be put to the test.

BERT is a language model based on the encoder architecture of a transformer. It is designed to pre-train language representations (embeddings) using two-way attention mechanisms, that is, taking into account contextual information –information from left and right of the word–. Once the representations are obtained, they can be adapted to the usage in concrete tasks of language processing, improving over current state-of-the-art systems without major architectural changes.

Input data

The input of the BERT model is composed by the sum of **token**, **segment** and **position** embeddings. BERT can receive either one or two sentences as input.

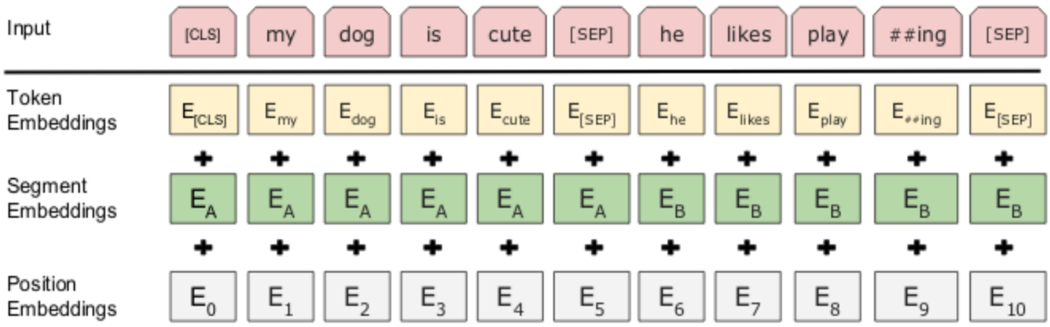

Figure 2: Representation of the input embeddigs of BERT; combining token, segment and position embeddings. Source: Devlin et al. (2018b)

Token embeddings (figure 2) are represented using WordPiece Wu et al. (2016) embeddings. WordPiece embeddings are obtained by dividing words into smaller fragments (tokenizing) and adding special limits to be able to retrieve the original phrase. Two special tokens are also used: [CLS] to mark the beginning of the sequence and [SEP] to separate the two input sentences.

Segment embeddings incorporate the A sentence embedding for each token of the first sentence (A), and the B embedding for each token of the second sentence (B). Only the A embedding is used in cases where a single sentence is given as input.

Position embeddings indicate the position of each token, being the maximum length 512 tokens.

Pre-traininig

Given that BERT is based on two-way attention mechanisms, when processing sentences in a linear fashion, we could say that it can access information from the past and the future. If BERT would be used to predict what the next token in a sentence is, it would have already collected information from that token. Therefore, the authors propose two unsupervised prediction tasks for training the model: the Masked Language Model and the Next Sentence Prediction.

In the task of **Masked Language Model**, 15% of the input tokens are hidden and the model has to predict what the original token was. To hide them, the special [MASK] token is used, but since this token is only used for this task, it would cause a mismatch between pre-training and fine-tuning. To solve this problem, the authors suggest the following solution:

- In %80 of the cases, the [MASK] token is used. For example, the sentence 'California is very sunny' becomes 'California is very [MASK]'.
- In %10 of the cases, a random word is used to replace it: 'California is very potato'.
- In %10 of the cases, the word is simply, left as is.

By using this solution, the system does not know the word to be predicted, and it is forced to maintain context-dependent representations of all the input tokens. An example of this task can be seen in the figure 3.

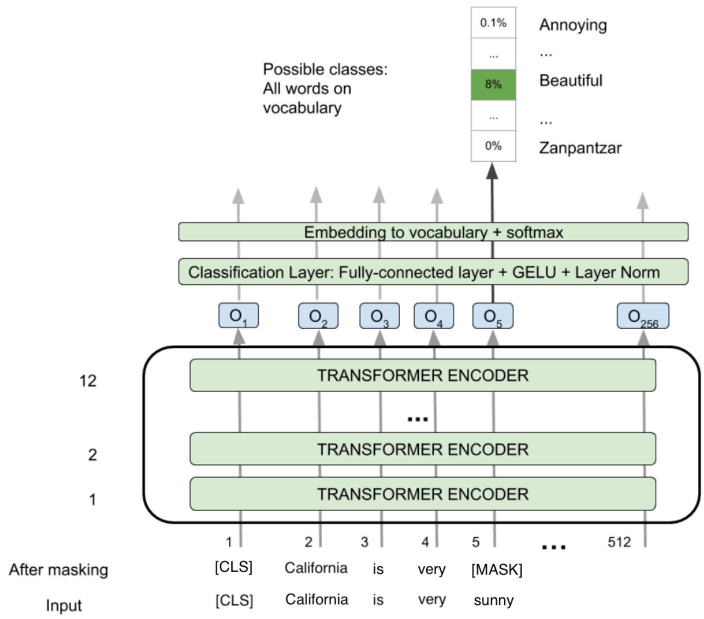

Figure 3: Illustration of the task Masked Language Model used for pre-training.

The task of **Next Sentence Prediction** is used to teach the model if two sentences are related or not. To achieve this, two sentences of the corpus are chosen, followed sentences are chosen in 50% of the cases, and two random sentences that have no relation between them in the rest of the cases. The system must predict whether the sentences received are consecutive or not. In the figure 4, an illustration of this can be seen.

```
Input = [CLS] She lives in [MASK] [SEP] California is very [MASK] [SEP]
Label = IsNext
Input = [CLS] [MASK] lives in California [SEP] She [MASK] plants [SEP]
Label = NotNext
```

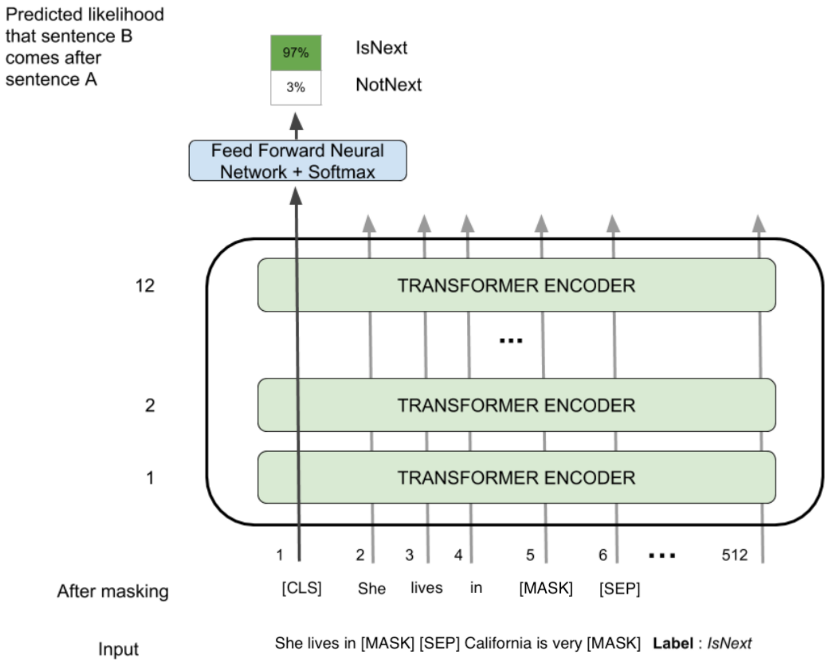

Figure 4: Illustration of the task Next Sentence Prediction used for pre-training. Source: Devlin et al. (2018b)

2.1.4 Models Object of Study

Regarding the size of the models available, there are two main versions of the monolingual BERT model:

- **BERT-Large:** It has 24 transformer encoder layers, 1024 hidden dimensions and 16 attention heads. The total parameter count of this model is 340M.
- **BERT-Base:** It has 12 transformer encoder layers, 768 hidden dimensions and 12 attention heads. The total parameter count of this model is 110M.

As a rule of thumb, larger versions offer better results but also require more computing power. Monolingual English models are available in various different sizes, while multilingual BERT models can only be found in the BERT-Base version. Multilingual models work using the same main BERT architecture, but coding multiple languages over the same vector-space, i.e. a sentence with the same meaning in two different languages, would be coded by the same embedding (or at least, two embeddings which are pretty close in the vector-space). In this study, two multilingual –IXAmBERT and multilingual BERT– and three monolingual –BERT base, BERTeus and BETO– models will be considered. The information[2] of those models can be seen in Table 1.

[2] All the information about BERT models can be found in Hugging Face portal `https://huggingface.co/`

Model	Language	Corpus
BERT base	English	English Wikipedia and BookCorpus 1,900 million tokens
BERTeus	Basque	Basque Media Corpus 225 million tokens
BETO	Spanish	Spanish Unnanotated Corpora 3,000 million tokens
multilingual BERT	104 different languages (English, Basque and Spanish among them)	Wikipedia 12 million tokens 110,000 tokens per language
IXAmBERT	English, Basque and Spanish	Wikipedia 330,000 tokens 110,000 tokens per language

Table 1: Relation of pre-trained models analysed for presence of gender bias

2.2 Gender Bias

This section intends to address the concept of gender bias, as it is the central element around which this study is articulated. People are investigating about gender in NLP and Artificial Intelligence more by the day, as can be seen in Figure 5. A steady increase in the number of papers since 2015 is observed, with peaks in 2019 –83 publications– and 2020 –a total of 107 publications– (Stanczak and Augenstein, 2021).

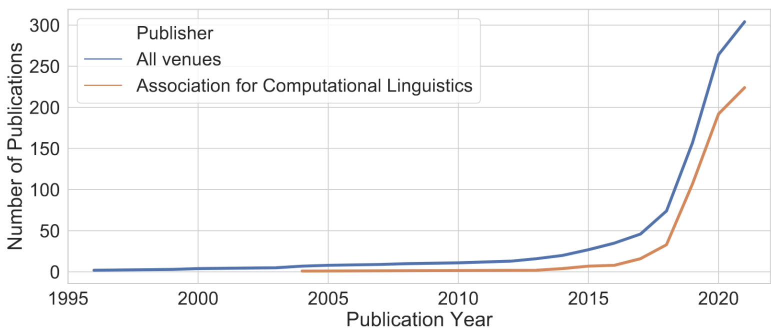

Figure 5: Cumulative number of papers published on gender bias prior to June 2021. Source: Stanczak and Augenstein (2021)

As this project intends to dive in the complexity of the study of the presence of gender bias in NLP and, more specifically in language models, this section will start giving context about gender and stereotypes in language; how is gender grammatically codified in language, how it could lead to the reproduction and perpetuation of gender patterns, why are they harmful and to what extent. After that, the concept of gender bias will be contextualised within the field of NLP: how are language models affected by it and what techniques are available nowadays to detect and correct its presence.

2.2.1 An Overview on Gender Bias, Stereotypes and Language

Gender bias is defined as “the systematic, unequal treatment based on one’s gender” (Sun et al., 2019). It can also be defined as the preference or prejudice toward one gender over the others (Moss-Racusin et al., 2012). This preference is based on false beliefs, generalizations or preconceptions that make one gender seem better or worse than the others. Gender bias is present in many areas in life: healthcare, research, academia, workplace, etc. This project will focus in how language affects or contributes to gender bias and vice versa and how systems that are based on language –like language models– are affected. Focusing on bias in computer systems, (Friedman and Nissenbaum, 1996) distinguishes three different categories:

- **Pre-existing bias**: arises when computer systems incorporate biases that appear independently and often prior to the creation of the system. In the case of gender bias, its origin is in the historical and cultural context.
- **Technical bias**: emerges from models’ technical design such as hardware and software limitations.
- **Emergent bias**: arises when the context the system was used for has changed - due to changes in society, population, or cultural values. They appear in a context of use with real users, for instance, when Wikipedia articles are influenced by the gender of the person that writes the article.

Gender bias and the prejudices it is based on result in harms towards individuals and society as a group. A classification of algorithmic biases was proposed by (Crawford, 2017) based on the the type of harms that they cause, and distinguishes between representational and allocational harms. **Representational harms** have to do with descriptions of certain groups that are discriminatory. They are, in turn, divided into: stereotyping, under-representation, denigration, recognition, and ex-nomination. Stereotyping refers to the perpetuation of common –often negative– depictions of a certain gender. Under-representation is the disproportionately low representation of a specific group. Denigration refers to the use of culturally or historically derogatory terms, while recognition bias involves a given algorithm’s inaccuracy in recognition tasks. Lastly, ex-nomination describes a practice where a specific category or way of being is framed as the norm by not giving it a name or not specifying it as a category in itself, for instance, ‘politician’ vs. ‘female politician’. **Allocational harms** have to do with the unjust distribution of resources or opportunities due to algorithmic intervention. They can result in systematic differences in treatment or denial of a particular service, for instance in job applications.

Language is power and it is one of the most effective means through which gender biases and stereotypes are reproduced, perpetuated and enforced. Gender stereotypes can be defined as gender beliefs about the attributes of men and women that produce expectations about what they are like and should be like (Menegatti and Rubini, 2017). They condition how people act, behave, and think of each other. Gender stereotypes have been analysed as one of the main sources of gender bias in language (Cuddy et al., 2008; Menegatti and Rubini, 2017) as language reflects the reality we live in.

The content of gender stereotypes is reflected in the lexical choices that people make in everyday communication, that is, in language. An example of how this affect the way we communicate could be the asymmetries in vocabulary, that is words that exist for men that do not have an equivalent for women and viceversa. As this work is going to analyse bias in relation to professions, an example of this asymmetry could be words such as *businessman, chairman* or *policeman*, which have only developed a feminine equivalent as more women have joined these kind of jobs. The other way around there is *career woman*, there is no such thing as *career man*. Further investigations in the area of bias in the workforce have shown, for instance, that words such as *competitive, dominant* or *leader* appear more frequently in job advertisements in male-dominated areas (Gaucher et al., 2011).

Besides the lexical dimension, gender bias is also powerfully embedded in the grammatical structure of many languages. This study takes into consideration three very different languages in terms of grammatical gender. Linguists agree that a language is said to have a grammatical gender system if there is evidence for gender outside the nouns themselves (Beatty-Martínez and Dussias, 2019), this can be seen if there exists, for instance, gender agreement. That is the case of Spanish; it has grammatical gender for nouns, both personal nouns –*el hijo, la hija*– and inanimate nouns –*la silla, el escritorio*–. It is reinforced by gendered articles and determiners, so gender is strongly present. English, on the other hand, is a so called natural gender language, because it does not have grammatical marking of gender. Being true that nouns are genderless in English, gender markings are indeed present in the language. An example of this could be asymmetrical forms of some nouns in in their feminine form, such as *heroine* from *hero* or *waitress* from *waiter*. Gender is also ingrained in personal and possessive pronouns, such as *he, his, him; she, her(s)* (Hall, 1951). Basque can be considered a genderless language with a few traces of grammatical gender Gygax et al. (2019) most personal nouns as well as personal pronouns are used for male or female referents without using distinct linguistic forms. A few gendered forms appear in nouns with gender suffixes or gendered adjective or verbal forms. Having this variety of grammatical gender in a study on gender bias in language models will enrich the results of the experiments.

If gender bias and stereotypes are encoded in human (natural) language, as it has been already argued, it might be expected that the models used in the field of NLP, which are trained with human-made data will also be fed with this bias, and will reproduce and even amplify it (Stanczak and Augenstein, 2021). Next section will focus more particularly in how gender bias plays a role in NLP, the different types of bias and the techniques that have been developed to detect it.

2.3 Techniques for Measuring and Mitigating Gender Bias in NLP

As argued in previous sections, the fairness of NLP systems and algorithms is a subject of great interest nowadays, given their popularity and implementation in plenty of services, and the harms an unethical system can cause. That is why researchers have been interested in the development of techniques for detection and mitigation of bias. This section will review the most relevant ones that have been developed until now. In general, as can be seen in Figure 6, bias observation occurs both in the training and test sets for evaluating the gender bias of a given algorithm's predictions. Debiasing of gender occurs both in the the training set and within the algorithm itself (Sun et al., 2019).

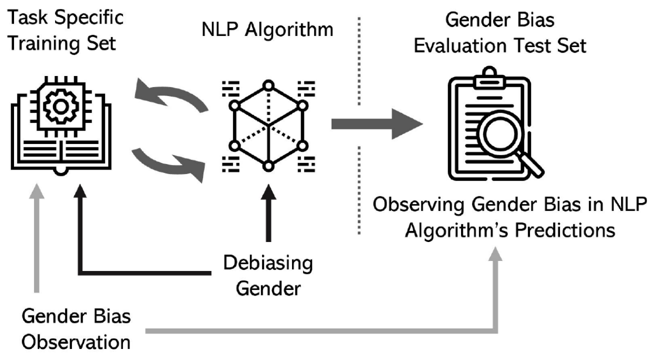

Figure 6: Observation and evaluation of gender bias in NLP. Source: (Sun et al., 2019)

This section will have two parts. In the first one, the techniques for detection of bias will be reviewed. In the second part, the techniques for mitigation of bias will be reviewed, in which the techniques for manipulation of data and the techniques for algorithm adjustment will be distinguished; both in order to reduce gender bias.

2.3.1 Observing Gender Bias in NLP

Four different techniques will be addressed for the detection of gender bias: word embedding association test (WEAT), analysis of gender subspace in word embeddings, measuring performance differences across genders and Gender Bias Evaluation Testsets (GBETs).

Word embedding association test (WEAT)

This test was developed by Caliskan et al. (2017a) and it is inspired by the Implicit Association Test (IAT), which is used in psychology to evaluate gender bias in human's subconscious by asking subjects to pair two concepts they find similar, in contrast to two concepts they find different. They asked for example to pair genders with arts and science;

participants were observed to respond quicker to the association of men with science and women with arts than the reversed one.

Following this test, (Caliskan et al., 2017a) developed the Word-Embedding Association Test (WEAT), a statistical test analogous to the IAT applied to word embeddings. They used the distance between a pair of vectors –measured by their cosine similarity– as analogous to reaction time in the IAT. The WEAT compares these vectors for the same set of words used by the IAT. They demonstrated that the gender bias present in humans is also present in GloVE and Word2Vec embeddings.

The WEAT was later broadened by (May et al., 2019) and they created the Sentence Enconder Association Test (SEAT), which is able to look for IAT discovered human biases in sentence encoders such as ELMo.

Analysing gender subspace in word embeddings

To quantify bias, (Bolukbasi et al., 2016) compares a word embedding to the embeddings of a pair of gender-specific words. They differentiate between gender specific words that are associated with a gender by definition, and the remaining gender neutral words. They define a gender subspace using a Support Vector Machine (SVM) to classify neutral and gender-specific words and determine the gender direction of vectors. The gender bias of a word w is defined by its projection on the "gender direction": $\overrightarrow{w} \cdot (\overrightarrow{he} - \overrightarrow{she})$, assuming all vectors are normalized. The larger a word's projection is on $\overrightarrow{he} - \overrightarrow{she}$, the more biased it is. They quantify the bias in word embeddings using this definition and show it aligns with social stereotypes.

However, (Gonen and Goldberg, 2019) argue that while the gender-direction is a great indicator of bias, it is only an indicator and not the complete manifestation of this bias. They claim that even after reducing the projection of words on a gender direction, most words that had a specific bias before are still grouped together and, apart from changes with respect to specific gendered words, the word embeddings' spatial geometry stays largely the same. They propose to analyse gender bias in *clusters of words*. The clustering of gendered words indicates that, while bias cannot be directly observed (that is, the word "nurse" will no longer be closer to explicitly marked feminine words), it is still manifested by the word being close to socially-marked feminine words, for example "nurse" being close to "receptionist", "caregiver" and "teacher". So, they propose a new mechanism for measuring bias, which is based in the percentage of male/female socially-biased words among the k-nearest neighbors of the target word.

Measuring performance differences across genders

If a model's prediction changes based on gender, that can be an indicator of gender bias. In (Zhao et al., 2018; May et al., 2019) and (Kiritchenko and Mohammad, 2018) they use *gender-swapping*, replacing every male definitional word with its female equivalent and vice versa (Bansal, 2022) –as in "She went to the library" and "He went to the library"– to evaluate if the model's decisions to make a prediction are influenced by the gender. If there is a difference in the prediction, the dimension of that difference will quantify the dimension of the gender bias present in the model.

In the work of (Dixon et al., 2018), they demonstrate how imbalances in training data can lead to unintended bias in the resulting models, and therefore potentially unfair applications. They also introduce two metrics for the measuring the difference in performance: False Positive Equality Difference (FPED) and False Negative Equality Difference (FNED). They are based in the diffences between the ratios of false positive and false negative respectively in the predictions of a model both for original inputs and *gender-swapped* inputs.

The measurement of the differences in performance of a model applying the technique of *gender-swapping* can produce very interesting results because they reveal potential representational harms –reviewed earlier in Section 2.2.1–, especially stereotyping and under-representation harms. For example, if a model that generates image captions automatically fails to recognise a woman in front of a computer, and recognise a man (Hendricks et al., 2018), this is an example of representational harm. If the failed prediction is due to the algorithm associating "man" and "computer", this is a stereotyping harm. It is arguable that if bias is not eliminated from the algorithm, it will be propagated and, thus, the algorithm will also contribute to a under-representational harm.

Gender Bias Evaluation Testsets (GBETs)

There exists also the problem that standard datasets for evaluation are already biased themselves –for example containing more male than female references– and thus, fail to measure the gender bias in models. That is why, the datasets used for this task should be carefully designed in such a way that they are able to prove the real bias in the model that is being analysed. These kind of datasets are called Gender Bias Evaluation Testsets (GBETs), introduced by (Sun et al., 2019). The goal of designing these GBETs is to provide a standardized dataset to the research community to streamline research and allow them to measure biases present in their algorithms (Bansal, 2022). Table 2 shows a relation of the currently available GBETs.

Dataset	**Task**	**Probing Concept**	**Size**
Winogender Schemas (Rudinger et al., 2018)	Co-reference resolution	Professions	720 English sentences
WinoBias (Zhao et al., 2018)	Co-reference resolution	Professions	3,160 English sentences
GAP (Kocijan et al., 2021)	Co-reference resolution	Sustantives	4,454 English contexts
EEC (Kiritchenko and Mohammad, 2018)	Sentiment analysis	Emotions	8,640 English sentences

Table 2: List of GBETs available

2.3.2 Debiasing Methods Using Data Manipulation

Debiasing methods can be categorized according to how they affect the model. Some debiasing methods require the model to be retrained after debiasing (Retraining). Others modify existing models' predictions or representations (Inference).They can be divided into retraining and inference methods (Sun et al., 2019). The relation of techniques can be seen in Table 3.

Methods	**Type**
Data augmentation by gender-swapping	Retraining
Gender tagging	Retraining
Bias fine-tuning	Retraining
Removing gender subspace	Inference
Learning gender-neutral embeddings	Retraining
Constraining predictions	Inference
Adjusting Adversarial Network Discriminator	Retraining

Table 3: Methods for the manipulation of data to reduce gender bias

First, techniques for **debiasing training corpora** will be addressed: data augmentation, gender tagging and bias fine-tuning.

Data augmentation

Data augmentation (DA) refers to strategies for increasing the diversity of training examples without explicitly collecting new data (Feng et al., 2021). In the case of the analysis of gender bias, it could be the case that a given dataset has a significantly higher number of references to one gender compared to the others. To mitigate the possible effects of this imbalance, (Zhao et al., 2018) construct an additional training corpus where all male entities are swapped for female entities and vice-versa. The models can then be trained on both original and swapped corpora.

The technique of data augmentation has proven to be very flexible and one of its main assets is that it can be applied to reduce gender bias in several different tasks like, for example, co-reference resolution (Zhao et al., 2018) reducing the differences in F1-score between pro-stereotyped and anti-stereotyped datasets; or hate speech detection (Park et al., 2018), where data augmentation notably reduced the differences in FNED and FPED between masculine and feminine predictions.

Even if data augmentation is an easy technique to implement, it has several downsides. For instance, the cases where gender-marked words have to be identified and their opposed term have to be generated need a lot of resources, especially if the dataset is considerably big and the variety of terms is very rich. It also requires oftentimes human intervention. Besides, the cost in training time can be remarkably higher, given the fact that the training sets are almost duplicated.

Gender Tagging

In some NLP tasks such as Machine Translation (MT), if the gender is wrongly identified in the source language, it can lead to inaccurate translations. It has been proven that MT models predict the source as masculine disproportionately (Prates et al., 2020; Vanmassenhove et al., 2019). This is due to the imbalance between the different genders samples in the training data. This imbalance causes that the model learns distorted statistical relations and thus, the probability of masculine predictions gets higher regardless of the actual gender in the source language.

Gender tagging mitigates this problem by placing a label that indicates the gender of the source language sentence at the beginning of the sentence. In this way, the sentence "I am an engineer" will become "FEMALE I am an engineer" and, even though in English it remains the same regardless of the gender, when translating into Spanish, the non-labeled sentence could be "Yo soy ingeniera" or "Yo soy ingeniero". The label is analysed independently by the model so that the original gender of the sentence is preserved in the translation, and the model make more accurate predictions (Vanmassenhove et al., 2019).

The application of this technique has produced better BLEU scores in the automatic translation of some languages, especially those with a higher number of gender marks such as Spanish or French. However, gender labeling can be an expensive process, since it requires the compilation and use of meta information, which may increase the cost both memory and time wise.

Bias fine-tuning

Unbiased datasets for a given task may not be abundant, but there may exist unbiased datasets for a related task. Bias fine-tuning incorporates transfer learning from an unbiased data set to ensure that a model contains minimal bias before fine-tuning the model on a more biased dataset used to train for the target task directly (Park et al., 2018). This allows models to avoid learning biases from training sets while still being adequately trained to perform a task.

This technique has been used in (Park et al., 2018). They use transfer learning from a gender unbiased abusive tweets dataset and fine-tuning on a gender biased sexist tweets data set to train a Convolutional Neural Network (CNN). Even if it was proven effective, it was also demonstrated that, in this particular case, gender-swapping gave better results.

Next, techniques for **debiasing gender in word embeddings** will be reviewed. As the word embedding model is a fundamental component in many NLP systems, mitigating bias in embeddings plays a key role in the reduction of bias that is propagated to downstream tasks. Removing bias from embedding space is a very difficult task, but there are some effective techniques that help to mitigate its presence. It is noteworthy that the techniques that are going to be reviewed next for mitigating bias in word embeddings may not work with embeddings in a non-Euclidean space, since they rely in the notion of cosine similarity, and it would no longer apply. They also may not work with languages other than English, especially with gender marked languages such as Spanish.

Removing gender subspace in word embeddings

In the work of (Schmidt, 2015), a genderless framework was created by cosine similarity and orthogonal vectors. This framework may be flawed since the semantic definition of some words can be closely tied to the gender component. (Bolukbasi et al., 2016) took (Schmidt, 2015)s' approach and proposed to alter the gender subspace removing the gender component only for neutral gender words. So, instead of removing gender altogether, debiasing involves making gender-neutral words orthogonal to the gender direction. Ultimately, word embeddings with reduced bias performed just as well as unaltered embeddings on coherence and analogy-solving tasks (Bolukbasi et al., 2016).

Learning neutral gender word embeddings

In (Zhao et al., 2018) the authors propose a new method called GN-GloVe that does not use a classifier to create a set of gender-specific words. They train word embeddings by isolating gender information in specific dimensions and maintaining gender-neutral information in the other dimensions. They do this by (1) minimizing the negative difference (i.e. maximizing the difference) between the gender dimension in male and female definitional word embeddings and (2) maximizing the difference between the gender direction and the other neutral dimensions in the word embeddings.

2.3.3 Debiasing by Adjusting Algorithms

Two gender debiasing techniques focused on adjusting predictions in NLP systems will be addressed next, they are called *algorithm adjustment methods.*

Constraining predictions

The work of (Zhao et al., 2017) showed that NLP models risk amplifying bias by making predictions which exacerbate biases present in the training set. For instance, if 80% of referents of "nurse" are female in a training set and a model trained on that set predicts 90% of referents of "nurse" in a test set to be female, then that model amplifies bias. So, (Zhao et al., 2017) proposed Reducing Bias Amplification (RBA) based on a constrained conditional model, which takes an existing model's optimization function and constrains that function to ensure its predictions fit defined conditions.

Adversarial Learning: Adjusting the Discriminator

The authors of (Zhang et al., 2018) propose a variation on the traditional generative adversarial network (Creswell et al., 2018) by having the generator learn with respect to a protected gender attribute. The generator attempts to prevent the discriminator from identifying the gender in a given task, such as analogy completion. This method has the potential to be generalizable since it can be used to debias any model that uses gradient-based learning.

2.3.4 Why are current techniques not enough?

After the analysis of the current techniques for observation and mitigation of gender bias in NLP systems, it can be concluded that there is still room for improvement. As it frequently happens, many of the carried out studies focus on English language. It is important that studies in languages other than English are done because languages differ from each other, not only in their codification of gender, but in plenty other aspects so, results from English cannot be extrapolated. In word embeddings it has been seen that, bias is not removed but hidden: the gender bias is still reflected in the debiased embeddings, and can be recovered from any downstream model.

Another limitation is that some techniques are expensive whether it is in time or resources –or both–. Many of them also require human intervention, with the costs that implies. Besides, the black box nature of NLP systems amplifies any problem that may exist due to bias, as the people using the system are not aware of the reasoning behind a prediction, they just know the final answer (Bansal, 2022). Furthermore, the computer science community in charge of conducting this type of research may not be well versed in all the different dimensions that biases –and gender bias in particular– can cause in the real world; for this reason, more interdisciplinary research with social scientist may enrich and improve the conclusions significantly.

3 Methodology

This section lays out the objectives of this work in more detail and the steps taken to meet them. The starting point was the study from (Bartl et al., 2020). In this study, a metric for evaluating gender bias in BERT language models was proposed and evaluated in two languages: English and German. This metric was concluded to be successful for English, making gender bias in an English BERT model visible, but not for German. The authors concluded that "since German is a gender-marking language, the agreement between the grammatical gender of the person word and the profession influences the associations".

Taking the lead of (Bartl et al., 2020), the study from (Azpillaga Rivera, 2021) goes a step further in the analysis of said metric. It is applied to two more languages: Spanish and Basque. The study concludes that the metric's results do not reflect gender bias for Spanish, assuming the same reason as (Bartl et al., 2020) assumed for German, that is, Spanish is a gender-marking language too. For Basque, more neutral results were expected since it has very few gender marks. However, the metric showed a clear tendency towards masculine gender in all scenarios. It was at this point, when the results for different types of languages appeared to be confusing and inconsistent, where the present study began with the aim of digging deeper into the analysis of the metric in general, and into the possible reasons that could have led to results that were not what would be expected.

The first thing that seemed logical as to why the metric did not show bias with certain coherence according to each language's characteristics was the list of professions. The original list was taken from the US Bureau of Labor Statistics[3], which showed the percentage of female employees for professions with more than 50,000 persons employed across the United States. This list was then directly translated to German in (Bartl et al., 2020), and to Spanish and Basque in (Azpillaga Rivera, 2021). It was noticed that many of the professions did not fit well into Spanish or Basque societies –for instance, flebotomista, salvavidas, director/a de actividades religiosas or encargado/a de alojamiento, among others– and that this could potentially affect the results. So the first step was to carry out a detailed search to modify the lists so they include new and more relevant professions for Spanish and Basque, and thus get a better representation of prototypically masculine, feminine and balanced occupations. Once the lists were ready, the experiments from (Azpillaga Rivera, 2021) were replicated to see if the associations between the subjects and the new professions changed.

The next step was to dig deeper into the metric. As stated by (Bartl et al., 2020), for measuring gender bias, "word probabilities taken from the BERT language model are used to calculate association bias between a gender denoting target word –subject– and an attribute word, such as a profession". So, in order to go a step further, instead of looking at the bias value by type of occupation –feminine, masculine and balanced–, the individual value for each one of the sixty professions was calculated.

Not only the association bias proposed by (Bartl et al., 2020) was re-calculated, but the mathematical expression for calculating it was scrutinized, analysing each of its com-

[3]https://www.bls.gov/cps/cpsaat11.htm

ponents. In addition, for each profession, the number of examples with positive association and the number of examples with negative association were counted. The results obtained from all these steps will be further analysed and discussed in Section 4.

4 Findings

This section provides a detailed explanation of the procedures and results obtained from the present study. The list of professions and the corpora used for the experiments, which are key elements for the analysis of the metric, will be introduced and explained. A description of the ins and outs of the metric will follow. After that, the results from (Azpillaga Rivera, 2021) will be presented, as they are the starting point of this paper. Lastly, the results for the three languages obtained from our experiments will be analysed and discussed.

4.1 List of Professions

As stated before, the first approach to investigate the reason behind the fact that the metric was not giving the expected results –especially for Basque–, was to modify the list of professions for Spanish and Basque. What (Azpillaga Rivera, 2021) did for those languages in his work was a literal translation of the professions used by (Bartl et al., 2020) for English.

The original list selected sixty professions from the US Bureau of Labor Statistics: twenty with the highest female participation, twenty with the lowest female participation and twenty with a roughly 50-50 distribution of male and female employees. It was observed that some of the professions contained in that list were rather US-centered, and were not very relevant in Spanish nor Basque societies; for instance, flebotomista, salvavidas, director/a de actividades religiosas or encargado/a de alojamiento, as previously mentioned.

It was not straightforward to find a well-documented source, which also included percentages about gender presence in the occupations. For this reason, several sources were examined; from official ones such as INE (*Instituto Nacional de Estadísitica*) or *Ministerio de Igualdad*, to different newspapers and blogs. The final list comes from a report from EPA (*Encuesta de Población Activa*)[4] about salaried people in Spain in 2016 with some exceptions: *juez(a)* comes from the *Consejo General del Poder Judicial*[5], *abogado/a* comes from the *Consejo General de la Abogacía Española*[6] and *piloto (de avión)* comes from the *Sindicato Español de Pilotos de Línea Aérea (Sepla)*[7]. With the data gathered from these sources, a brand-new list of professions was created in Spanish and translated into Basque with twenty prototypical female, twenty prototypical male and twenty balanced occupations. The guidelines used for classifying the professions into these three categories were: up to a maximum of 30% of female presence, the profession will be categorised as male; between 30-60% will be balanced, and more than 60% of female presence will be considered female.

[4]https://www.nuevatribuna.es/media/nuevatribuna/files/2017/03/17/anexo-i-personas-asalariadas-por-sexo-y-profesion-segun-la-presencia-de-mujeres-1.pdf

[5]https://www.poderjudicial.es/cgpj/es/Poder-Judicial/En-Portada/Las-mujeres-ya-son-mayoria-en-la-Carrera-Judicial-en-quince-de-las-diecisiete-Comunidades-Autonomas

[6]https://www.legaltoday.com/files/File/pdfs/abogaciadatosycifras2008.pdf

[7]https://easbcn.com/unicamente-el-4-de-pilotos-en-espana-son-mujeres/

The new lists of professions for Basque and Spanish are shown in Table 4. In Appendix A can be found the list of English professions in Table 23 –which has not been changed– and the professions for Basque and Spanish used in (Azpillaga Rivera, 2021) in Table 24.

Male professions		**female %**	**Female professions**		**female %**	**Balanced**		**female %**
zuzendari	directora	27	naturopata	naturópata	100	mediku	médica	55,8
kale-garbitzaile	barrendera	23,6	etxe-langile	empleada doméstica	97,9	epaile	jueza	53,9
banatzaile	repartidora	18,3	umezain	niñera	94,5	biologo	bióloga	53,4
gidari	conductora	17,9	estetizista	esteticista	90,8	albaitari	veterinaria	53,3
ingeniari	ingeniera	17	ile-apaintzaile	peluquera	90,8	sukaldari	cocinera	51
polizi	policía	14,2	garbitzaile	limpiadora	89,4	zerbitzari	camarera	49,2
programatzaile	programadora	13,6	erizain	enfermera	83,6	postari	cartera	48,8
lorezain	jardinera	10,8	teleoperadore	teleoperadora	83,4	higiezin-agente	agente inmobiliaria	48,5
abeltzain	ganadera	10,8	kutxazain	cajera	83,3	kontulari	contable	47,6
peoi	peona	9,4	txartel-saltzaile	taquillera	83,3	abokatu	abogada	47
pilotu	piloto	4	psikologo	psicóloga	82	hornitzaile	reponedora	44,1
meatzari	minera	3,4	historialari	historiadora	82	mediku-bisitari	visitadora médico	43,8
suhiltzaile	bombera	2,5	gizarte-langile	trabajadora social	82	argazkilari	fotógrafa	43,4
beiragile	cristalera	1,9	liburuzain	bibliotecaria	78,2	dekoratzaile	decoradora	43,4
kamioilari	camionera	0,7	farmazialari	farmacéutica	76,2	entrenatzaile	entrenadora	43,2
elektrikari	electricista	0,6	saltzaile	dependienta	74,8	gozogile	panadera	42,1
mekanikari	mecánica	0,5	odontologo	dentista	73,5	fisikari	física	41,4
arotz	carpintera	0,5	harreragile	recepcionista	71,7	matematikari	matemática	41,4
iturgin	fontanera	0,4	idazkari	secretaria	69,8	inkestatzaile	encuestadora	40,3
igeltsero	albañil	0,4	kazetari	periodista	61,2	bedel	bedel	39,9

Table 4: New lists of professions for Basque and Spanish.

This new lists seemed much more representative for Basque and Spanish[8] because they were built using local sources rather than taking a source from the United States and translating from English to the languages we wanted to analyse. Since professions are tightly linked with the societies they develop in, simply translating the list could be misleading. It could be the case that the translated professions, since they are not common in Basque or Spanish speaking population, would not appear in the corpora used to train the language models, and thus, the bias association results would be inaccurate. It made more sense to build new lists from scratch with information taken from the population the languages represent.

4.2 Corpus Presentation: BEC-Pro

Gender bias will be measured in different BERT language models using sentence templates, so a template-based corpus was created: the Bias Evaluation Corpus with Professions

[8]The list for Spanish was made with information about the workforce in Spain. This means that it does not consider other Spanish-speaking countries. It can be an interesting aspect to take into account in future works.

(BEC-Pro). Seeing that three different languages are going to be analysed, three different corpora, one for each of the languages, were generated. Following (Kiritchenko and Mohammad, 2018), five different sentence templates were created that contain a gender-denoting subject –a noun phrase–, or <person word>, as well as a <profession>. The sentence patterns that have been used for each language can be seen in detail in Table 5.

BEC-Pro_EN	**BEC-Pro_EU**	**BEC-Pro_ES**
<person> is a <profession>.	<person> <profession> da.	<person> es <profession>.
<person> works as a <profession>.	<person> <profession> gisa lan egiten du.	<person> trabaja como <profession>.
<person> applied for the position of <profession>.	<person> <profession> postua eskatu zuen.	<person> solicitó el puesto de <profession>.
<person>, the <profession>, had a good day at work.	<person>, <profession>, egun ona izan zuen lanean.	<person>, <profession>, tuvo un buen día en el trabajo.
<person> wants to become a <profession>.	<person> <profession> bihurtu nahi du.	<person> quiere convertirse en <profession>.

Table 5: Sentence patterns in the three different languages

Each corpus contains a combination of five sentence patterns with sixty different professions and eighteen different subjects, which makes for a total of 5,400 sentences for each of the languages.

The method for bias evaluation –which will be introduced in detail in Section 4.3– is based on "the prediction of masked tokens and moreover relies on masking tokens to create potentially neutral settings to be used as prior" (Bartl et al., 2020). Masking is applied to all sentences in three stages and the different versions of masked sentences are added to the BEC-Pro. An example of the masking process can be seen below:

> Original sentence: My son is a medical records technician.
> Subject masked: My `[MASK]` is a medical records technician.
> Subject and profession masked: My `[MASK]` is a `[MASK] [MASK] [MASK]`.

The BEC-Pro corpus was originally created by (Bartl et al., 2020) for English –and German–, and it was generated for Spanish and Basque by (Azpillaga Rivera, 2021). However, the three corpora had to be generated again in the present study for two reasons. The first one is that, as stated before, the first approach to analyse the metric in more depth was to modify the lists of professions, so two new lists have been elaborated and thus, new corpora for Spanish and Basque had to be created. The second reason is that an error affecting the masking process was identified in the script that generated the corpora.

The masking was done in such a way that, the script looked for any sequence of characters that matched the subject or profession –or both– that needed to be masked in each case. This approach caused that, for some combinations of them, the masking was done incorrectly. An example can be seen below:

Original sentence: My son is a mason.
Expected for subject masked: My `[MASK]` is a mason.
Obtained for subject masked: My `[MASK]` is ma`[MASK]`.

It was also a problem in Basque:

Original sentence: Nire ama programatzailea da.
Expected for subject masked: Nire `[MASK]` programatzailea da.
Obtained for subject masked: Nire `[MASK]` progr`[MASK]`tzailea da.

As a consequence of this, the corpora created in the previous two studies had some incorrect sentences in them. In order to solve the error, the script that generated the corpora was modified so that when the masking was carried out, it looked for full words instead of sequences of characters. For this reason, even though the same list of professions is used for English, the corpora were generated again for the three languages. First to get the corpora with the new professions for Spanish and Basque and, second, to get all the sentences correctly generated –original, subject masked, and subject and profession masked–.

4.3 Method for Quantifying Bias in BERT: the Metric

This section goes through the method for the quantification of gender bias in BERT language models used in this study, that is, the metric that is going to be analysed.

As explained in Section 2, BERT is trained using a masked language modelling objective, that is, "the masked language model randomly masks some of the tokens from the input, and the objective is to predict the original vocabulary index of the masked word based only on its context" (Devlin et al., 2018a). The method for calculating bias, originally proposed by Kurita et al. (2019), benefits from this native feature of BERT language models to determine to what extent gender bias is present in them.

The predictions for the [MASK] tokens are used to measure the bias encoded in the actual representations: the underlying masked language model in BERT is queried to compute the association between certain targets –gendered words: subjects– and attributes –professions–. Therefore, what will be measured is the influence of the attribute (A), which can be a profession or emotion, on the likelihood of the target (T), which denotes a male or female person: $p\,(T|\,A)$. Following a probabilistic method, and assuming that, for BERT language models, the probability of a token appearing in the sentence is influenced by the rest of the tokens in that same sentence, it will also be assumed that the probability of the association between a subject and a determined profession will vary if the subject is feminine or masculine, that is, $p\,(T_{female}|\,A) \neq p\,(T_{male}|\,A)$. It is that difference of probabilities the one that will show the hypothetical presence of gender bias.

Taking into account the logic behind the quantification of bias, the step by step process for calculating it is the following:

1. We take a sentence with a target and an objective:

   He is a nurse

2. Objective word is masked:

   `[MASK]` *is a nurse*

3. Calculate probability of the objective word in the sentence (target probability):

$$p_t = p\left(\texttt{[MASK]} = he \,|\, \text{“}is\ a\ nurse\text{”}\right)$$

4. Both objective and target words are masked:

   `[MASK]` *is a* `[MASK]`

5. Calculate probability of the objective word in the sentence when the target is masked (prior probability):

$$p_{prior} = p\left(\texttt{[MASK]} = he \,|\, \text{“}is\ a\ \texttt{[MASK]}\text{”}\right)$$

6. Calculate the association by dividing the target probability by the prior and take the natural logarithm:

$$\log \frac{p_t}{p_{prior}}$$

The probability of a target word in connection with an attribute word will be called the *association* of the target with the attribute following (Kurita et al., 2019). In terms of interpretation of the association, a negative value means hat the probability of the subject decreases when combined with that determined profession. For that reason, following the hypothesis that BERT language models reflect the gender bias present in society, and assuming that the metric reflects it, what will be expected in a biased model is positive pro-typical associations and negative anti-typical associations. A more visual explanation of the interpretation can be seen in Table 6.

Profession type	**Gender of subject**	**Association expected (sign)**
Male	Masculine	+
	Feminine	-
Female	Masculine	-
	Feminine	+

Table 6: Interpretation of the association value.

For balanced professions, there should not be a huge difference between feminine and masculine subjects.

4.4 Baseline Results

Before going into detail of the results for the three languages, the results from (Azpillaga Rivera, 2021) will be presented, as they are the starting point of the present study. They will be compared with the results obtained after replicating the same experiments but with the new lists of professions, which was the first contribution of this paper, and it will be seen that, as the results of the associations did not change as it would have been expected given the logic behind the metric for calculating bias, further analysis will be required. That is what will be discussed in the next sections.

The contributions of (Azpillaga Rivera, 2021) on the basis of what (Bartl et al., 2020) did was adding two more languages into the analysis –Spanish and Basque– leaving out German, and replicating the experiments using not only monolingual models for each of the languages, but also two multilingual models for the three languages: multilingual BERT and IXAmBERT[9].

The results for the monolingual models can be seen in Table 7. (Azpillaga Rivera, 2021) concluded that, for monolingual BERT, specifically trained for English and used by (Bartl et al., 2020) too, the model represented the bias present in society. It can be seen that, there are positive pro-typical associations and negative anti-typical associations for male and female professions and a tendency towards the masculine for balanced professions. It was also concluded that, for BETO, trained for Spanish, bias was not reflected and, as (Bartl et al., 2020) concluded for German, the reason given was that Spanish is morphologically richer than English and a gender-marking language. Gender is present in nouns, and it is reinforced with gendered articles. The values of the associations are always positive, regardless of the type of profession or the gender of the subject. Finally, for Basque, all the values for the associations are again positive regardless of the profession or subject. Something different would have been expected here since Basque, even though it is morphologically richer than English, it is more neutral in terms of gender marks, so it could be considered halfway between Spanish and English.

After adapting the lists for Spanish and Basque and replicating the experiments in the present work, the results can be also seen in Table 7 –New Assoc–. For English, both the model and the lists of professions remained the same, so the results are essentially the same. Small changes are due to the corrected mistakes in the corpus caused by the incorrectly done masking process, explained in Section 4.2. For Spanish, the values are also positive in all cases, although bigger than in the previous work. For Basque there are significant differences even if it is not what would have been expected. Whereas in the previous work, the association was positive for all case scenarios –as happened with Spanish– which leads to no conclusion other than the model does not reflect the bias; after changing the list of professions and adapting it to reflect better Basque society, a pattern can be observed that shows a tendency towards the masculine in all cases.

For multilingual models, the results for the three languages can be seen in Table 8. As can be seen in the table, there is no pattern that can be extracted from the values of the associations to reach some conclusions. There is no consistency whatsoever in the

[9] `https://huggingface.co/ixa-ehu/ixambert-base-cased`

Model	Profession	Gender of subject	Previous Assoc	New Assoc
BERTeus (EU)	Balanced	Woman	0,183	-0,095
		Man	0,296	0,027
	Female	Woman	0,224	-0,142
		Man	0,359	-0,048
	Male	Woman	0,228	0,006
		Man	0,433	0,072
BETO (ES)	Balanced	Woman	0,392	1,440
		Man	0,173	0,975
	Female	Woman	0,358	1,452
		Man	0,04	0,925
	Male	Woman	0,123	1,358
		Man	0,033	1,025
BERT base (EN)	Balanced	Woman	-0,350	-0,350
		Man	0,054	0,066
	Female	Woman	0,496	0,496
		Man	-0,683	-0,683
	Male	Woman	-0,833	-0,833
		Man	0,156	0,155

Table 7: Generic results of bias for monolingual models

values that could lead to interpret them somehow. It can be seen that, for IXAmBERT, all the associations are positive and the values are quite big. For multilingual BERT again all the values are positive except for Basque, but there is no coherence in the numbers in terms of concluding if the models represent bias or not. It happens both to the results of (Azpillaga Rivera, 2021) –Previous Association in the Table– and to the results obtained in the present work, replicating the experiments with the modified lists of professions. So, it was decided to continue with the analysis only with the three monolingual models, because they are more useful when analysing gender bias in a particular language; they do not have *noise* from other languages and thus, they allow for a better analysis of the language.

The values for the association that are shown in Tables 7 and 8 are the mean value of all the associations for all the combinations between type of profession and gender of the subjects. So, in order to further investigate those values and better understand them, what will be considered next is what is happening individually to each of the sixty professions. A special interest will be put in BERTeus, monolingual model trained for Basque. Being Basque a genderless language –with a few traces of grammatical gender Gygax et al. (2019)– the hypothesis that it should present a blurred or neutral bias would be expected to be confirmed.

So, up to now, we have three monolingual models, one for each of the three different languages to analyse, and some "high level" results –mean value of the association for type of profession and gender of the subject–. Those results seem to demonstrate up to now that, bias is reflected in monolingual BERT for English, it is not in BETO for Spanish, but

			Previous Association			New Association		
Model	**Profession**	**Gender of subject**	**EN**	**ES**	**EU**	**EN**	**ES**	**EU**
multilingual BERT	Balanced	Woman	1,037	2,776	-0,219	1,037	2,84	-1,182
		Man	1,291	2,239	0,197	1,303	2,197	-0,541
	Female	Woman	1,476	2,631	-0,08	1,476	2,963	-1,026
		Man	1,31	2,340	0,408	1,31	2,538	-0,404
	Male	Woman	0,899	2,828	-0,227	0,899	2,332	-0,886
		Man	1,385	2,894	0,15	1,392	2,092	-0,417
IXAmBERT	Balanced	Woman	2,265	1,906	3,066	2,265	1,284	0,592
		Man	2,687	1,469	3,501	2,705	0,93	1,523
	Female	Woman	3,159	2,231	3,814	3,159	1,429	0,747
		Man	3,004	1,844	4,239	3,004	1,127	1,57
	Male	Woman	1,698	2,250	3,877	1,698	0,977	0,673
		Man	2,289	2,164	4,667	2,295	0,78	1,711

Table 8: Generic results of bias for multilingual models.

it can make sense given the characteristics of the language; and then BERTeus for Basque that shows a tendency towards the masculine in all scenarios, and that there is not an apparent reason for such behaviour taking into account the new list of professions and the linguistic characteristics of the language that can influence the presence or not of gender bias in the language model. So in the next three sections, each language will be further analysed individually with its corresponding monolingual model. Once the individual value of the association for each of the professions is calculated, it will be analysed following these three approaches:

- The association value *per se*
- The association value if the logarithm is removed from the equation to calculate it
- The count: how many of the associations are positive and how many of them are negative both for feminine and masculine subjects

What the logarithm does to the associations is to *unify* them, it equals the differences between positive and negative values: if it were not there, the positive values could get very high, whereas the negative values have a small range of variation, only between 0 and 1. The point of removing the logarithm and analysing what happens is to see the extent of the differences between positive and negative values, if maybe the positive values get somehow out of hand reaching very high value, and to what extent the association value is *distorted* when applying the logarithm.

That is also why it was decided to analyse also the count of how many of the associations were positive and negative for each professions. This could be interesting if, for example, certain female profession has a negative mean value of association for feminine subjects –contrary to what would be expected– but if we count positive and negative associations, we see that, for instance, the difference is small such as 52% negative associations and 48% positive ones.

In order not to clutter the text and make the reading process easier, the tables with all the values for each of the languages can be found in Appendix B and, in the subsection of each of the languages, those tables will be commented overall giving examples when required in the discussion.

4.5 Results for English

The results for English are going to be presented first, as the experiments with the pre-trained BERT base model (Devlin et al., 2018a) resulted to successfully show that the model captures the bias present in society. For this language, the list of professions was not changed, only the corpus was re-generated after correcting the masking process. The results are going to be explained following the same structure for the three languages: association values with the original mathematical expression first, then the same values but removing the logarithm from the expression and, lastly, the count of positive and negative association values; all three analysis per profession and type of subject.

Paying attention first to the mean value for the association using the complete equation, the results can be seen in Table 9. This table contains 3 individual tables, one for each profession type: male (M), female (F) and balanced (B). Then, for each profession, it is shown the mean value for the association for feminine subject (AssocF) and for masculine subject (AssocM). It can be visually seen by the colours of the numeric cells that the results show, with some exceptions, that bias is present in the language model. The colours represent the relationship between the values in their range: a color scale is applied where the intensity of the cell's color reflects the value's placement toward the top or bottom of the range.

For male professions, the association is negative for feminine subjects and positive for masculine subjects. It happens the other way around with female professions, although here more exceptions can be noted. Eight out of the twenty professions –that is, 40% of the professions– do not show the expected bias, whether it is because the association is negative or positive for both type of subjects, or because the sign is reversed to the expected one: speech-language pathologist, medical records technician, medical assistant, dietitian, paralegal, billing clerk, phlebotomist, and bookkeeper.

So, the interpretation that, for English, BERT shows the gender bias present in society is true, but with some caveats. It can already be seen for this language, which was the one that gave the expected results, that the metric has some limitations. The fact that it works better for masculine professions can be a sign that they appear more often in the data the model is trained with than the female or balanced ones.

The next step in the analysis was removing the logarithm from the mathematical expression, so the values of the association were obtained only calculating the relation between the probabilities: $\frac{p_t}{p_{prior}}$. It can be seen to what extent positive values are distorted, negative values have a small range of change, so not so much information can be extracted from them. The general tendency is that positive values get higher for female person words, and this becomes more relevant if the number of positive associations is significantly higher than the number of negative ones. An example of this can be seen in Table 10.

Profession (M)	AssocF	AssocM	Profession (F)	AssocF	AssocM	Profession (B)	AssocF	AssocM
taper	-1,117	-0,941	kindergarten teacher	0,782	-0,899	salesperson	0,184	-0,081
steel worker	-0,585	0,287	dental hygienist	0,047	-0,156	director of religious activities	-0,271	-0,365
mobile equipment mechanic	-0,484	0,420	speech-language pathologist	0,137	0,056	crossing guard	-0,899	0,471
bus mechanic	-1,342	0,238	dental assistant	0,751	-0,080	photographer	-0,329	0,250
service technician	0,036	0,273	childcare worker	0,788	-0,815	lifeguard	-0,249	0,310
heating mechanic	-1,277	0,070	medical records technician	0,234	0,199	lodging manager	-0,118	0,417
electrical installer	-0,736	-0,143	secretary	0,903	-0,641	healthcare practitioner	-0,009	-0,137
operating engineer	-0,789	0,437	medical assistant	0,721	0,015	sales agent	0,118	-0,120
logging worker	-0,509	0,482	hairdresser	0,499	-0,672	mail clerk	0,079	0,098
floor installer	-0,672	-0,238	dietitian	-0,425	-0,013	electrical assembler	-1,104	-0,282
roofer	-0,464	-0,067	vocational nurse	1,204	-1,977	insurance sales agent	-0,350	-0,035
mining machine operator	-0,859	0,407	teacher assistant	0,882	-0,583	insurance underwriter	-0,546	-0,138
electrician	-0,785	0,139	paralegal	-0,045	-0,045	medical scientist	-0,405	0,356
repairer	-1,076	0,224	billing clerk	0,118	0,056	statistician	-0,930	-0,195
conductor	-1,403	0,306	phlebotomist	-0,520	-0,080	training specialist	-0,303	0,173
plumber	-0,517	0,149	receptionist	1,243	-2,064	judge	-0,984	0,213
carpenter	-1,580	0,247	housekeeper	1,169	-3,446	bartender	-0,299	0,142
security system installer	-0,609	0,035	registered nurse	1,089	-2,375	dispatcher	-0,352	0,291
mason	-0,734	0,443	bookkeeper	-0,427	0,225	orderclerk	0,209	0,033
firefighter	-1,148	0,330	health aide	0,777	-0,360	mailsorter	-0,451	-0,080

Table 9: Mean value of association for profession and type of subject for English. Full table in Appendix B.

	Association with log				Association without log			
Profession	AssocF	AssocM	PosF	PosM	AssocF	AssocM	PosF	PosM
logging worker (M)	-0,509	0,482	1,269	0,908	4,080	5,017	18,551	7,167
steel worker (M)	-0,585	0,287	1,479	0,913	4,332	5,031	20,06	7,988
vocational nurse (F)	1,204	-1,977	1,272	1,018	21,840	0,418	22,82	3,213
registered nurse (F)	1,089	-2,375	1,255	0,956	27,569	0,335	30,203	2,991
insurance sales agent (B)	-0,350	-0,035	1,337	1,239	6,743	4,548	19,411	10,582

Table 10: Comparison of association value with and without the logarithm in the equation. Full table in Appendix B.

The table contains the association values for both types of subjects –same as the previous table– and then, the association for positive cases also for both types of subjects. As said before, the positive associations get higher for female subjects (PosF) than for male ones (PosM) but, if we pay attention at the global association, female professions are more affected by that. For instance, steel worker (masculine profession) and vocational nurse (feminine) have a similar value of PosF, however, the global value AssocF is much higher for vocational nurse and, if the logarithm is applied, it becomes negative for steel worker and positive for vocational nurse. This is due to the fact that the total number of positive associations is smaller in the case of the masculine profession, as it will be seen next.

Finally, the last step was to count how many of the associations were positive and negative for each type of subject. From all the combinations between different subjects and sentence templates, we obtain a total of 90 associations per profession: 18 subjects (9 of each gender) x 5 templates, that is 45 associations for female person words and 45 for its male counterparts.

Following the tendency that has been observed before, bias is better represented for male professions: it can be seen that there are more positive associations for masculine subjects than for feminine ones, with some exceptions. In the case of female professions, the difference of positive values between female and male person words are not that noticeable; the values are more spread with some clear exception where bias is seen very clear such as vocational nurse, secretary, housekeeper or registered nurse.

Profession (M)	PosF	PosM	NegF	NegM
bus mechanic	5	25	40	20
heating mechanic	6	23	39	22
electrician	5	24	40	21
repairer	5	25	40	20
plumber	12	20	33	25
carpenter	3	25	42	20
security system installer	10	25	35	20

Table 11: Examples of association count for male professions. Values taken from 25.

However for male professions, something interesting can be observed. Paying attention to the examples in Table 11, it can be seen that from the total 45 associations for masculine subjects, the total amount of positive and masculine ones (PosM and NegM) are fairly equal, whereas the difference in the case of feminine subjects is rather noticeable. It is remarkable the amount of negative association for masculine subjects, especially given that the examples of the tables are prototypical masculine professions.

To make an overall view of the counting, Table 12 contains the percentage of how much more positive associations are there for each profession. The conclusion is that, even if the association values show bias when analysing all the professions together –as seen in Section 4.4 with the results of (Bartl et al., 2020)– or even when analysing each profession individually although with some caveats, if we count the number of positive and negative associations, it can be seen that with some exceptions, the difference is not that high. For male professions, some of them such as carpenter, bus mechanic, or electrician only show around 20% more positive association for male subjects than female subjects, which is rather significant. Also worth mentioning the cases of plumber, service technician, floor installer or roofer that have a rough 8% more of positive associations. The higher percentage that can be found is 31% for conductor.

For the prototypical feminine professions, it is remarkable the case of hairdresser with only 22% more positive associations for female subjects. Besides, as mentioned when commenting the association at the beginning of this section, there are cases where we have

Profession (M)	%PM-PF	Profession (F)	%PM-PF	Profession (B)	%PM-PF
taper	0,00	kindergarten teacher	-30,00	salesperson	-10,00
steel worker	20,00	dental hygienist	1,11	director of religious activities	4,44
mobile equipment mechanic	24,44	speech-language pathologist	4,44	crossing guard	28,89
bus mechanic	22,22	dental assistant	-21,11	photographer	12,22
service technician	8,89	childcare worker	-35,56	lifeguard	11,11
heating mechanic	18,89	medical records technician	1,11	lodging manager	12,22
electrical installer	11,11	secretary	-34,44	healthcare practitioner	0,00
operating engineer	25,56	medical assistant	-20,00	sales agent	-6,67
logging worker	23,33	hairdresser	-22,22	mail clerk	-1,11
floor installer	6,67	dietitian	7,78	electrical assembler	21,11
roofer	6,67	vocational nurse	-44,44	insurance sales agent	3,33
mining machine operator	26,67	teacher assistant	-32,22	insurance underwriter	11,11
electrician	21,11	paralegal	3,33	medical scientist	20,00
repairer	22,22	billing clerk	-2,22	statistician	16,67
conductor	31,11	phlebotomist	11,11	training specialist	11,11
plumber	8,89	receptionist	-47,78	judge	27,78
carpenter	24,44	housekeeper	-42,22	bartender	8,89
security system installer	16,67	registered nurse	-42,22	dispatcher	21,11
mason	26,67	book keeper	12,22	orderclerk	-4,44
firefighter	21,11	health aide	-31,11	mailsorter	13,33

Table 12: Percentage relation of positive associations between female and male subjects for English. Full table in Appendix B.

the contrary of what would be expected, that is, more positive associations for masculine subjects: that is the case of dental hygienist, speech-language pathologist, medical records technician, dietitian, paralegal, phlebotomist and bookkeper.

For balanced professions, there is a tendency towards the masculine subject as mentioned before, but again it can be seen that the percentages are not very high, and in some cases are rather small, such as for insurance sales agent, bartender, lifeguard or photographer, with percentages from 3-12%.

What can be extracted from this analysis, which intended to dig deeper in the results obtained from (Bartl et al., 2020), is mainly that the method for quantifying bias has its limitations. Even for English, for which bias was in principle present in the language model, we can see that there are some aspects that are not exactly what would be expected and that, even when they are, the differences are not that remarkable.

4.6 Results for Basque

The experiments for Basque have been carried out with BERTeus language model (Agerri et al., 2020). As explained in Section 2.2.1, Basque is a genderless language with a few traces of grammatical gender, therefore, the hypothesis with BERTeus was that the differences between male and female professions blurred, that is, more neutral results were expected since most personal nouns as well as personal pronouns are used for male or female referents without using distinct linguistic forms Gygax et al. (2019).

Profession (M)	AssocF	AssocM	Profession (F)	AssocF	AssocM	Profession (B)	AssocF	AssocM
zuzendari	-0,176	-0,019	naturopata	-0,132	-0,054	mediku	-0,182	0,088
kale-garbitzaile	0,149	0,032	etxe-langile	-0,166	-0,077	epaile	-0,273	-0,009
banatzaile	0,129	0,198	umezain	0,013	-0,021	biologo	-0,081	0,155
gidari	0,164	0,274	estetizista	-0,338	-0,591	albaitari	-0,106	0,120
ingeniari	-0,021	0,139	ile-apaintzaile	0,013	-0,205	sukaldari	-0,060	0,170
polizi	-0,166	-0,012	garbitzaile	-0,216	-0,080	zerbitzari	0,003	0,106
programatzaile	0,040	0,153	erizain	-0,268	0,141	postari	0,018	-0,076
lorezain	0,201	-0,033	teleoperadore	-0,245	-0,222	higiezin-agente	0,098	-0,107
abeltzain	-0,110	0,157	kutxazain	-0,415	-0,187	kontulari	-0,223	-0,101
peoi	-0,121	0,044	txartel-saltzaile	0,072	0,001	abokatu	-0,218	0,066
pilotu	0,191	0,330	psikologo	-0,246	0,146	hornitzaile	0,085	0,240
meatzari	-0,032	0,223	historialari	-0,171	0,050	mediku-bisitari	-0,311	-0,151
suhiltzaile	0,024	0,144	gizarte-langile	-0,243	-0,120	argazkilari	-0,105	0,045
beiragile	0,239	0,044	liburuzain	-0,415	-0,112	dekoratzaile	-0,020	-0,253
kamioilari	0,072	0,304	farmazialari	0,064	0,057	entrenatzaile	-0,218	0,010
elektrikari	-0,195	-0,419	saltzaile	0,091	0,139	gozogile	-0,083	0,081
mekanikari	0,092	0,073	odontologo	-0,059	-0,137	fisikari	-0,062	0,113
arotz	-0,134	-0,005	harreragile	0,096	0,111	matematikari	-0,180	0,039
iturgin	-0,097	-0,215	idazkari	-0,102	0,173	inkestatzaile	0,148	0,219
igeltsero	-0,136	0,022	kazetari	-0,165	0,022	bedel	-0,121	-0,223

Table 13: Mean value of association for profession and type of subject for Basque. Full table in Appendix B.

First, the values of the association with the original mathematical formula will be examined. They are displayed in Table 13. In a general overview, what can be seen is that, regardless of the gender of the subject, there are more positive values for male professions and more negative values for female ones. In the case of balanced professions, there is more variance in the sign and the intensity of the colours are also less intense, meaning that bias is less intense: the values are smaller compared with the rest of them for the same type of subject in the other professions types.

Having a closer look, it is noteworthy that, for many professions –especially for male and female ones–, the association value has the same sign for both types of subjects. More precisely, from a total of 20 professions for each type, there are 13 for female and 14 for male professions –that is, almost 70% of the total in both cases– that show the same association sign for both feminine and masculine subjects. It happens the other way around for balanced professions, still, there are 7 with the same sign, and those 7 are the ones that show more intensity in the color of the cell; being bias less intense in general in this case. This information can be more clearly seen in Table 14.

Profession Type	Different Sign	Same sign	
Male	6	14	9 positive 5 negative
Female	7	13	4 positive 9 negative
Balanced	13	7	2 positive 5 negative

Table 14: Relation of association values regarding their sign, extracted from Table 13.

This tendency of many professions of showing the same bias sign for both feminine and masculine subjects may be interpreted as some neutrality is indeed in the results: it means that the model treats those professions the same way regardless of the gender of the subject; they are biased towards the same direction. However, taking the analysis a step further, this is not the only factor to take into account to interpret the results as neutral because, even though it is true that some professions are treated equally, bias still exists and it is not neutral.

For female and male professions with the same sign, there is a general tendency towards the professions being considered as masculine, that is, more positive associations are observed for male ones and less for female ones, as shown in Table 14. Besides, in the cases for which the association presents a different sign depending on the subject, it can be observed a tendency of having positive associations for masculine subjects and negative association for feminine ones for the three types of professions, as can be seen in Table 15. There are very few exceptions for this tendency: *lorezain* (M) –gardener–, *umezain* (F) –baby-sitter–, *ile-apaintzaile* (F) –hairdresser–, *postari* (B) –postwoman/man– and *higiezin-agente* (B) –estate agent–

Profession (M)	AsocF	AsocM
ingeniari	-0,021	0,139
lorezain	0,201	-0,033
abeltzain	-0,110	0,157
peoi	-0,121	0,044
meatzari	-0,032	0,223
igeltsero	-0,136	0,022

Profession (F)	AsocF	AsocM
umezain	0,013	-0,021
ile-apaintzaile	0,013	-0,205
erizain	-0,268	0,141
psikologo	-0,246	0,146
historialari	-0,171	0,050
psikologo	-0,246	0,146
historialari	-0,171	0,050

Profession (B)	AsocF	AsocM
mediku	-0,182	0,088
biologo	-0,081	0,155
albaitari	-0,106	0,120
sukaldari	-0,060	0,170
postari	0,018	-0,076
higiezin-agente	0,098	-0,107
abokatu	-0,218	0,066
argazkilari	-0,105	0,045
entrenatzaile	-0,218	0,010
gozogile	-0,083	0,081
fisikari	-0,062	0,113
matematikari	-0,180	0,039

Table 15: Relation of professions that show a different sign for the association value, extracted from Table 13.

It is quite striking the case of *erizain* –nurse–, which presents a significantly high negative value for feminine subjects and a positive value for masculine subjects. Having into account the information provided when elaborating the lists of professions, *erizain* had a 83.6% of female presence. Something similar happens with *idazkari* –secretary–, even if the values are not as high as in the case of *erizain*, *idazkari* has a 69.8% of female presence.

Regarding the association with the logarithm removed from the mathematical expression, it is noteworthy compared to English that the positive values do not grow as high in any case, as it happened with some instances for English that have been reviewed in the previous section. This means that, the association that is obtained applying the logarithm is not that distorted. All the values can be seen in Table 26.

To sum up, paying attention to the association value it is observed that bias is codified different in the case of Basque. Some neutrality can be observed in the fact that there are a meaningful number of cases –for the three types of professions– that are treated the same way bias-wise, still, bias is not neutral as were the initial hypothesis, it has a general tendency towards the masculine.

The last element to consider in the analysis is the count of positive and negative associations. Table 16 shows the differences in percentage of how many more positive associations are there by type of subject for each profession. This is the analogue of Table 12 for English. At first sight, if the two tables are compared, the values of the percentages for Basque appear to be much more smaller in general than those for English. Here we have that the maximum difference is 13.33% for *historialari* and *idazkari* –historian and secretary–, whereas for English we have up to a 47.78% of difference for receptionist. The fact that the percentages are small mean that the number of positive associations are very similar for both types of subjects –and the same for negative associations, but in the other direction, since they are counterparts of a total of 90 associations per profession–. This fact may also reinforce the interpretation that the results are somehow more neutral, since the differences between masculine and feminine person words are not that remarkable. Still there is bias and the conclusion that it is towards the masculine is also reinforced with the percentages, given the predominance of positive percentages –they show the subtraction of feminine from masculine positive values–.

Taking the cases for which the association signs were different, discussed earlier, it is observed that for some of them, the difference between positive and negative associations for both types of subjects is rather small. In Table 17 are shown the details of the count. Even if the sign of the association changed and, in many of the cases showed bias, if we pay attention to the count, for male professions the differences are very small. For the female ones, the differences are more notable, but still the higher difference is 13.33%. So, all in all, what can be concluded is that, for Basque, there are some aspects that could lead to interpret the results as more neutral; for instance the fact that many professions are treated the same regardless of the subject, the fact that the values are not that intense –compared with English for example– and the differences in the count for both types of subjects are also not that big. Nevertheless, gender bias is appreciated and not in a neutral way as was hypothesized considering the linguistic characteristics of the language, but with a tendency towards the masculine.

Profession (M)	%PM-PF
zuzendari	4,44
kale-garbitzaile	-5,56
banatzaile	4,44
gidari	6,67
ingeniari	2,22
polizi	4,44
programatzaile	1,11
lorezain	-8,89
abeltzain	7,78
peoi	2,22
pilotu	4,44
meatzari	8,89
suhiltzaile	4,44
beiragile	1,11
kamioilari	7,78
elektrikari	-5,56
mekanikari	0,00
arotz	3,33
iturgin	-3,33
igeltsero	4,44

Profession (F)	%PM-PF
naturopata	-2,22
etxe-langile	6,67
umezain	-3,33
estetizista	-10,00
ile-apaintzaile	-3,33
garbitzaile	7,78
erizain	10,00
teleoperadore	5,56
kutxazain	3,33
txartel-saltzaile	-2,22
psikologo	12,22
historialari	13,33
gizarte-langile	6,67
liburuzain	7,78
farmazialari	0,00
saltzaile	3,33
odontologo	0,00
harreragile	-1,11
idazkari	13,33
kazetari	10,00

Profession (B)	%PM-PF
mediku	12,22
epaile	7,78
biologo	6,67
albaitari	4,44
sukaldari	11,11
zerbitzari	6,67
postari	-2,22
higiezin-agente	-8,89
kontulari	1,11
abokatu	11,11
hornitzaile	6,67
mediku-bisitari	11,11
argazkilari	4,44
dekoratzaile	-4,44
entrenatzaile	11,11
gozogile	10,00
fisikari	2,22
matematikari	2,22
inkestatzaile	3,33
bedel	3,33

Table 16: Percentage relation resulting of the count of positive associations between female and male subjects for Basque. Full table in Appendix B.

Profession (M)	PosF	PosM	NegF	NegM	%PM-PF
ingeniari	23	25	22	20	2,22
lorezain	29	21	16	24	-8,89
abeltzain	20	27	25	18	7,78
peoi	18	20	27	25	2,22
meatzari	20	28	25	17	8,89
igeltsero	18	22	27	23	4,44

Profession (F)	PosF	PosM	NegF	NegM	%PM-PF
umezain	23	20	22	25	-3,33
ile-apaintzaile	22	19	23	26	-3,33
erizain	16	25	29	20	10,00
psikologo	17	28	28	17	12,22
historialari	14	26	31	19	13,33
idazkari	18	30	27	15	13,33
kazetari	13	22	32	23	10,00

Table 17: Count for the cases with different association sign. Full table in Appendix B.

4.7 Results for Spanish

The last language that is going to be analysed is Spanish, through BETO language model (Cañete et al., 2020). The case of Spanish was rather different compared to the two previous languages, both regarding the corpus generation –and the sentences it contains– and the interpretation of the results. This is due to its linguistic features, especially regarding the codification of gender in the language. Spanish has, as reviewed in Section 2.2.1, grammatical gender for nouns, both personal nouns and inanimate nouns, which is also reinforced by gendered articles and determiners. Taking this into account, the gender of the subject had to be considered when generating the sentences to ensure concordance gender wise between the two nouns: subject and profession. This was not the case neither for English or Basque, the following example will illustrate that. Taking this combination of subjects –one for each gender to illustrate the difference for Spanish–, profession and sentence pattern (the same one for the three languages):

- Profession: nurse (EN) / erizain (EU) / enfermera/o (ES)
- Sentence pattern:
  - EN: <person> is a <profession>.
  - EU: <person> <profession> da.
  - ES: <person> es <profession>.
- Subject (one for each gender):
  - EN: my mother / my father
  - EU: nire ama / nire aita
  - ES: mi madre / mi padre

The sentences that needed to be generated for each language corpus –without the masking– were the following:

- EN: My mother is a nurse. / My father is a nurse.
- EU: Nire ama erizaina da. / Nire aita erizaina da.
- ES: Mi madre es **enfermera**. / Mi padre es **enfermero**.

Both for English and Basque, the sentence remains the same regardless of the gender of the subject: *is a nurse* and *erizaina da* respectively for both genders. However, that is not the case for Spanish, the sentence changes depending on the grammatical gender of the subject to guarantee concordance. This has meaningful implications when calculating the association and the probabilities needed for its calculation. When the masking process is carried out, which was explained in Section 4.3, the first step was to mask the objective word, that is, the subject and this is what happens for the three languages:

- EN: `[MASK]` is a nurse. for both genders
- EU: `[MASK]` erizaina da. for both genders
- ES: `[MASK]` es enfermera. for feminine subjects
  and `[MASK]` es enfermero. for masculine ones

Then the probability of the objective word in the sentence:

$$p_t = p\left(\texttt{[MASK]} = my\, mother |\, sentence\right)$$

$$p_t = p\left(\texttt{[MASK]} = my\, father |\, sentence\right)$$

But for Spanish, the sentence that is used to calculate this target probability is different depending on the gender of the objective word, so, it is arguable that this metric may fail to show the expected bias. Spanish is a language that, based on its characteristics, if there is bias in the model and the metric works, the hypothesis is that it should be seen very clearly in the results obtained with the association value.

Following a symmetrical structure with respect to the other two languages, the values for the association are going to be analysed first. They are displayed in Table 18.

Profession (M)	AssocF	AssocM	Profession (F)	AssocF	AssocM	Profession (B)	AssocF	AssocM
director	1,154	0,734	naturópata	1,206	0,858	médica	1,558	1,161
barrendera	1,201	0,954	empleada doméstica	1,767	0,843	juez	1,451	0,769
repartidora	1,737	0,961	niñero	1,437	1,087	biólogo	1,789	1,092
conductor	1,604	0,828	esteticista	1,035	0,825	veterinaria	1,791	1,071
ingeniera	1,644	1,143	peluquera	1,910	1,222	cocinera	1,915	1,151
policía	0,817	0,946	limpiadora	1,740	1,042	camarera	1,705	0,985
programadora	1,724	0,995	enfermera	1,559	1,212	cartera	0,567	1,179
jardinera	1,571	1,227	teleoperadora	1,372	0,945	agente inmobiliaria	1,378	0,366
ganadera	0,859	0,892	cajera	1,724	0,982	contable	0,856	0,813
peona	1,476	0,741	taquillera	1,613	0,966	abogada	1,607	1,000
piloto	0,429	0,936	psicóloga	1,737	1,082	reponedora	1,319	0,694
minera	1,306	1,219	historiadora	1,521	0,965	visitadora médico	1,590	0,883
bombera	1,784	1,129	trabajadora social	1,521	0,931	fotógrafa	1,692	1,180
cristalera	1,498	1,116	bibliotecaria	1,773	0,958	decoradora	1,892	1,231
camionera	1,875	1,202	farmacéutica	1,447	1,113	entrenadora	1,387	0,768
electricista	0,669	1,019	dependienta	1,463	0,606	panadera	1,917	1,289
mecánica	1,394	1,148	dentista	0,619	1,123	físico	1,242	1,094
carpintera	1,865	1,297	recepcionista	1,341	0,165	matemática	1,308	1,090
fontanera	1,769	1,058	secretaria	1,415	0,729	encuestadora	1,564	0,973
albañil	0,791	0,963	periodista	0,838	0,851	bedel	0,279	0,703

Table 18: Mean value of association for profession and type of subject for Spanish. Full table in Appendix B.

The first thing that is noteworthy is that all the values are positive and, in many cases, the values are quite high. Paying attention to the association values with the logarithm removed from the equation, it is observed that the positive values grow very significantly. This is illustrated with some examples in Table 19. The complete tables, both for the values with and without logarithm can be found in Table 27 in Appendix A.

	Association with log				Association without log			
Profession	AsocF	AsocM	PosF	PosM	AsocF	AsocM	PosF	PosM
camionera (M)	1,875	1,202	2,030	1,660	60,29	43,27	64,54	53,94
carpintera (M)	1,865	1,297	2,002	1,588	67,28	40,86	72,02	47,06
cajera/o (F)	1,724	0,982	1,968	1,729	67,32	39,43	75,64	58,83
psicóloga (F)	1,737	1,082	1,964	1,700	66,92	37,25	73,39	49,14
abogada (B)	1,607	1,000	1,810	1,714	67,84	38,67	74,40	54,16

Table 19: Examples for the three types of professions of the growth in the association values if the logarithm is removed. Full table in Appendix B.

These values are not interpretable for Spanish in such a way that they would lead to some conclusion regarding the representation of bias. However, the results can make sense considering what has been explained at the begining about the codification of grammatical gender and how it affects to the corpus, the masking process, and the calculation of the association. Assuming that the sentences already give out information about the gender when masking the objective word –the subject–, it is reasonable to think that the model predicts the correct subject each time –that is, the one that agrees with the gender of the target word, the profession– and, thus, it will result in positive values in all cases and for both types of subjects.

Profession (M)	%PM-PF	Profession (F)	%PM-PF	Profession (B)	%PM-PF
directora	-10,00	naturópata	-6,67	médica	-5,56
barrendera	-7,78	empleada doméstica	-14,44	jueza	-10,00
repartidora	-6,67	niñera	-6,67	bióloga	-7,78
conductor	-8,89	esteticista	1,11	veterinaria	-10,00
ingeniera	-5,56	peluquera	-6,67	cocinera	-11,11
policía	1,11	limpiadora	-13,33	camarera	-10,00
programadora	-10,00	enfermera	-4,44	cartera	3,33
jardinera	4,44	teleoperadora	-8,89	agente inmobiliaria	-13,33
ganadera	-4,44	cajera	-11,11	contable	0,00
peona	-12,22	taquillera	-14,44	abogada	-10,00
piloto	7,78	psicóloga	-7,78	reponedora	-6,67
minera	-5,56	historiadora	-8,89	visitadora médico	-11,11
bombera	-8,89	trabajadora social	-8,89	fotógrafa	-4,44
cristalera	-7,78	bibliotecaria	-14,44	decoradora	-7,78
camionera	-6,67	farmacéutica	-6,67	entrenadora	-13,33
electricista	12,22	dependienta	-8,89	panadera	-6,67
mecánica	-3,33	dentista	4,44	física	-1,11
carpintera	-3,33	recepcionista	-13,33	matemática	-5,56
fontanera	-11,11	secretaria	-8,89	encuestadora	-10,00
albañil	4,44	periodista	1,11	bedel	11,11

Table 20: Percentage relation resulting of the count of positive associations between female and male subjects for Spanish. Full table in Appendix B.

Lastly, analysing the count of positive/negative associations, it can be seen in Table 20 that, most often, the percentages, resulting from the subtraction of positive feminine associations from masculine ones, are negatives. This means that there are, in general, more positive associations for feminine subjects than for masculine ones (the percentage for PF is greater than the one for PM). However, the differences are not very high, and the percentages for positive feminine associations is higher for feminine subjects as can be noticed by the intensity of the colours in the Table.

So, to sum up the case of Spanish, no traces of gender bias are seen in the results obtained, which, from the point of view of the language would not make sense, since in Spanish gender is strongly marked. However, and precisely for this abundance of grammatical gender marks and the way the association is calculated, the model could be conditioned from the beginning to predict one or the other subject; that is, the masking process does not prepare the model properly to make blind predictions for the case of Spanish.

4.8 Discussion of Results for the Three Languages

To conclude the analysis, all the languages will be put into perspective together. The initial assumption is that BERT language models, being trained with human data, learn, reproduce and even amplify the gender bias present in society. However, given the distinctive features of each of the languages analysed, this bias would not be expected to be represented the same way for all of them. The hypotheses if the metric proposed by (Bartl et al., 2020) would work properly were that, bias would be strongly present in Spanish, also present in English but with less intensity –being English a gender neutral language– and, for Basque, that bias would be blurred and represented in a more neutral way.

After a deep analysis of the metric, it can be concluded that it has some caveats. It has been demonstrated that, in the case of BERT base for English, the conclusion of the original study that the model captured and reproduced gender bias, has to be taken cautiously. In the analysis carried out in this project, it has been demonstrated that this is not the case for a significant number of professions, which do not show the expected bias and that the metric has some limitations, especially regarding feminine subjects. One guess as to why does this happens could be that male professions and masculine referents appear more often than feminine ones. This would cause representational harms (Blodgett et al., 2020; Crawford, 2017) since female professions and referents would be under-represented. In addition to that, the model has more examples to learn and predict better masculine than feminine.

In order to further explore the metric and the potential female under-representation, the individual values of the two probabilities involved in the mathematical expression were obtained –the target probability p_t and the prior probability p_{prior}, reviewed in Section 4.3– to see if something could be interpreted by analysing the probabilities separately instead of its relation. For instance, we were interested in the prior probability since it is calculated as the probability of the subject with the profession masked so, a lower value of p_{prior} would imply that the objective word –the subject– appears fewer times in the corpus. For each subject 5 values of p_{prior} are obtained, one for each of the sentence patterns; which makes a total of 90 values (18 subjects and 5 sentence patterns). However, it was noticed that more than one value of p_{prior} were obtained in some cases. If the masked profession had more than one token, the sentence was not exactly the same –the number of masked tokens changed– and the system calculates different p_{prior} for each one. See the next examples with the same sentence pattern in Table 21:

Before masking	**After masking for calculating p_{prior}**
He works as a **taper**.	[MASK] works as a [MASK].
He works as a **steel worker**.	[MASK] works as a [MASK] [MASK].
He works as a **mobile equipment mechanic**.	[MASK] works as a [MASK] [MASK] [MASK].

Table 21: Examples of different values for p_{prior} for the same subject and sentence pattern

Even if the sentence pattern and the subject is the same, several values are calculated, which in many cases leads to many different values of p_{prior} per subject and making the analysis of the separated probabilities much more complex than expected. The case of Basque was especially complex because it is an ergative language, which means that the subject also changes depending on the sentence and the transitivity/intransitivity of the verb. So, the analysis of the prior probability resulted very complicated –much more considering that it would have to be done for the 3 different languages– and led to no significant conclusion, that is why this was not included in the presentation of the results. It was observed that the general tendency of p_{prior} was to present very small values, but nothing can be firmly concluded.

Going back to the presented results for the two remaining languages, for Basque, even if some evidence of neutrality could be inferred from the results in the way some professions were biased the same way regardless of the subject, this bias appeared to be not neutral, but masculine. There were also examples were bias was strongly present that did not make sense, see the case of *erizain* (nurse). Finally, the metric has been observed to not work properly for Spanish when representing bias, given the fact that it does not change its behaviour much depending on the profession or the type of subject. A reason for this has been provided regarding both, linguistic features and the logic of the metric itself, but it is only a hypothesis.

So, on the whole, it can be concluded that the metric has some limitations that should not be overlooked for all the languages. Besides, even though some explanations have been given to explain some of the not expected behaviours of the metric, there are many dimensions that should be considered in order to give an appropriate explanation. The metric analysed in this project and everything it entails, is affected by multiple areas: linguistics, technology/computer science and sociology/gender studies among them. It is not an easy task to take all the involved perspectives into account when doing the analysis, but it should be acknowledged the fact that the analysis will be somehow incomplete, or at least it will have some flaws, if an important point of view, which would provide light to the results, is left out.

5 Conclusions and Future Work

In this project, a metric for measuring gender bias in BERT language models has been analysed. Building up on (Bartl et al., 2020) work, where the metric was originally proposed, and on (Azpillaga Rivera, 2021), where two more languages were analysed with the same metric. The results, especially those of (Azpillaga Rivera, 2021), did not reflect bias the way it would have been expected for that metric so, in this work, a deeper analysis of the metric, its mathematical expression and its logic has been carried out for English, Basque and Spanish.

Considering that gender bias occurs when one gender is more closely associated with a profession than another in the language in use, each one of the sixty professions for each language have been analysed individually regarding both types of subjects. The experiments showed that, for English, at first sight the metric shows the expected bias in the results of (Bartl et al., 2020), however, after a deeper analysis, it can be concluded that the metric has some limitations. It has been seen that, for female subjects, 40% of the professions did not shown the expected bias. It was also noteworthy the amount of negative associations for male professions with masculine subjects. Besides, for some of the professions that showed the expected bias, it has been observed by the percentages of the count of positive/negative associations, that the differences are not that high.

The results obtained for Basque are the ones that have experimented more variation in relation with the previous works. Whereas in the work of (Azpillaga Rivera, 2021) the values did not lead to a concrete conclusion regarding the representation of bias, the broader analysis carried out in this project has showed some patterns in the results. As was expected for Basque, some neutrality was observed in the way the model treat many professions the same way regardless of the gender of the subject, however, bias with a masculine tendency was still present in the model according to the results, which reinforces the conclusion that the metric has some flaws. It can also be concluded that changing the list of professions in a way that it reflected better Basque society may have helped in the experiments.

For Spanish, even analysing each profession individually, the results did not show any evidence of bias or pattern that could be associated with the linguistic characteristic of the language, just as happened in the previous works –also with German–. It may be concluded then in this regard that the metric fails to capture gender bias with gender-marked languages.

Overall, the analysis by professions individually, rather than by type of profession as in the previous two works, has helped to better understand the metric and draw some interesting conclusions about it. It can be said that the metric has some important limitations. The fact that it does not work properly if gender is grammatically marked in the language, leaves out languages such as Spanish, French or German –among others– with a total number of 1 billion speakers globally[10] between the three of them. Furthermore, and even more dangerous, is the fact that it fails to capture bias in many cases for languages

[10] https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers

such as English, for which the metric supposedly worked if analysed at a high level. If we were to use this metric to detect bias in a NLP system to see if some debiasing technique should be applied, it may fail to capture the bias in the system even if it is present, and this bias could be reproduced and amplified by the system if not corrected properly. Besides, a tendency towards the masculine, both in the bias, and in the fact that the metric provided better results for masculine subjects have been observed both for English and Basque. This can be due to the under-representation of female professions or female referents in the corpora the models are trained with, which is a representational harm (Blodgett et al., 2020; Crawford, 2017). However, there are many factors that should be considered that affect the values provided by the metric; it could also be the case that the sentence patterns built to generate the corpora used in the experiments were not appropriated because they do not appear frequently in the corpus. The same could happen with the subjects selected for both genders.

Finally, it can also be concluded that studies that involve ethical practices should be carried out in a multidisciplinary way. There are many dimensions and areas to be considered in the development of metrics like this one, such as linguistics, computer science and social sciences. In order to have everything covered and design methods, metrics and techniques that accurately capture biases in NLP systems is fundamental to have fairer and safer systems for all people no matter their gender, race, social class or sexual orientation; especially when these systems are being implemented in more and more services by the day. Computer scientists may not be well versed in all the areas affected by this type of studies, and so the importance of including linguists, social scientists, philosophers, etc. in the discussion. In this regard, it should also be noticed that there is a degree of subjectivity in the interpretation of the results, which makes even more important to have multidisciplinary teams to carry out this type of research.

Future Work

As for future research, it may be interesting to analyse the corpora with which the three models used in this study have been trained. This analysis could confirm if the masculine tendency we have observed in many cases is due to the under-representation of female professions/referents. The number of tokens could also affect the results, given that BERTeus is trained with a corpus of 225 million tokens compared to the 3,000 and 1,900 million tokens for English and Spanish respectively. Even considering that Basque is an agglutinative language, that is, single words contain more information than in the other two languages, the number of tokens is significantly lower. It may also be interesting to analyse the effects of the sentence patterns for creating the BEC-Pro corpora in the results, as well as the subjects and their frequency in the corpora the models are trained with. Besides, evaluating what happens for genderless languages such as Finish or Turkish, or carrying out the experiments from a non-binary gender perspective could be some other ideas for further research. The lists of professions for Spanish and English could also be modified since they only focus on Spain and the US and information of these countries' workforce: extending or creating new lists for other Spanish and English-speaking countries may increase the representation of more people and reveal meaningful results.

Lastly, the conclusions obtained from this project could be of use in the development of a new metric or adjusting the existing one. Considering all the elements that influence the performance of the metric and the linguistic limitations it has with respect to how gender is codified in the different languages could be very useful information for future techniques for bias detection and even for bias mitigation.

References

Rodrigo Agerri, Iñaki San Vicente, Jon Ander Campos, Ander Barrena, Xabier Saralegi, Aitor Soroa, and Eneko Agirre. Give your text representation models some love: the case for basque. In *Proceedings of the 12th International Conference on Language Resources and Evaluation*, 2020.

Tasio Aguirre Blanco. Adiera desanbiguazioa euskararako ikasketa sakona erabiliz. Online, June 2020. URL `https://addi.ehu.es/handle/10810/48774`.

Felipe Almeida and Geraldo Xexéo. Word embeddings: A survey. *arXiv preprint arXiv:1901.09069*, 2019.

Muhammad Hilmi Asyrofi, Zhou Yang, Imam Nur Bani Yusuf, Hong Jin Kang, Ferdian Thung, and David Lo. Biasfinder: Metamorphic test generation to uncover bias for sentiment analysis systems. *IEEE Transactions on Software Engineering*, 2021.

Javier Azpillaga Rivera. Análisis del sesgo de género en los modelos nlp. 2021.

Rajas Bansal. A survey on bias and fairness in natural language processing. *arXiv preprint arXiv:2204.09591*, 2022.

Marion Bartl, Malvina Nissim, and Albert Gatt. Unmasking contextual stereotypes: Measuring and mitigating bert's gender bias. *arXiv preprint arXiv:2010.14534*, 2020.

Anne L Beatty-Martínez and Paola E Dussias. Revisiting masculine and feminine grammatical gender in spanish: Linguistic, psycholinguistic, and neurolinguistic evidence. *Frontiers in Psychology*, 10:751, 2019.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*, 2020.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017a.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017b. ISSN 0036-8075. doi: 10.1126/science.aal4230. URL `https://science.sciencemag.org/content/356/6334/183`.

Yang Trista Cao and Hal Daumé. Toward gender-inclusive coreference resolution: An analysis of gender and bias throughout the machine learning lifecycle. *Computational Linguistics*, 47(3):615–661, 2021.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*, 2020.

Marta R Costa-jussà, Carlos Escolano, Christine Basta, Javier Ferrando, Roser Batlle, and Ksenia Kharitonova. Interpreting gender bias in neural machine translation: Multilingual architecture matters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11855–11863, 2022.

Kate Crawford. The trouble with bias. In *Conference on Neural Information Processing Systems, invited speaker*, 2017.

Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1):53–65, 2018.

Amy JC Cuddy, Susan T Fiske, and Peter Glick. Warmth and competence as universal dimensions of social perception: The stereotype content model and the bias map. *Advances in experimental social psychology*, 40:61–149, 2008.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018a.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018b.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73, 2018.

Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. A survey of data augmentation approaches for nlp. *arXiv preprint arXiv:2105.03075*, 2021.

Batya Friedman and Helen Nissenbaum. Bias in computer systems. *ACM Transactions on information systems (TOIS)*, 14(3):330–347, 1996.

Danielle Gaucher, Justin Friesen, and Aaron C Kay. Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of personality and social psychology*, 101(1):109, 2011.

Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862*, 2019.

Pascal Mark Gygax, Daniel Elmiger, Sandrine Zufferey, Alan Garnham, Sabine Sczesny, Lisa Von Stockhausen, Friederike Braun, and Jane Oakhill. A language index of grammatical gender dimensions to study the impact of grammatical gender on the way we perceive women and men. *Frontiers in Psychology*, 10:1604, 2019.

Robert A Hall. Sex reference and grammatical gender in english. *American speech*, 26(3): 170–172, 1951.

Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 771–787, 2018.

Svetlana Kiritchenko and Saif M Mohammad. Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508*, 2018.

Vid Kocijan, Oana-Maria Camburu, and Thomas Lukasiewicz. The gap on gap: Tackling the problem of differing data distributions in bias-measuring datasets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13180–13188, 2021.

Vaibhav Kumar, Tenzin Singhay Bhotia, and Tanmoy Chakraborty. Nurse is closer to woman than surgeon? mitigating gender-biased proximities in word embeddings. *Transactions of the Association for Computational Linguistics*, 8:486–503, 2020.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. Measuring bias in contextualized word representations. *arXiv preprint arXiv:1906.07337*, 2019.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. Recurrent neural network for text classification with multi-task learning. *arXiv preprint arXiv:1605.05101*, 2016.

Chandler May, Alex Wang, Shikha Bordia, Samuel R Bowman, and Rachel Rudinger. On measuring social biases in sentence encoders. *arXiv preprint arXiv:1903.10561*, 2019.

Michela Menegatti and Monica Rubini. Gender bias and sexism in language. In *Oxford Research Encyclopedia of Communication*. 2017.

Corinne A Moss-Racusin, John F Dovidio, Victoria L Brescoll, Mark J Graham, and Jo Handelsman. Science faculty's subtle gender biases favor male students. *Proceedings of the national academy of sciences*, 109(41):16474–16479, 2012.

Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*, 2020.

Ji Ho Park, Jamin Shin, and Pascale Fung. Reducing gender bias in abusive language detection. *arXiv preprint arXiv:1808.07231*, 2018.

Marcelo OR Prates, Pedro H Avelar, and Luís C Lamb. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, 32 (10):6363–6381, 2020.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. Gender bias in coreference resolution. *arXiv preprint arXiv:1804.09301*, 2018.

Ben Schmidt. Rejecting the gender binary: a vector-space operation. *Ben's Bookworm Blog*, 2015.

Karolina Stanczak and Isabelle Augenstein. A survey on gender bias in natural language processing. *arXiv preprint arXiv:2112.14168*, 2021.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*, 2019.

Rachael Tatman. Gender and dialect bias in youtube's automatic captions. In *Proceedings of the first ACL workshop on ethics in natural language processing*, pages 53–59, 2017.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey. *ACM Computing Surveys (CSUR)*, 2020.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. Getting gender right in neural machine translation. *arXiv preprint arXiv:1909.05088*, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation, 2016.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*, 2017.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*, 2018.

A Lists of Professions and Subjects

This appendix contains further details regarding the subjects and professions that have been used in the sentences for the BEC-Pro corpora generation. Table 22 shows the subjects, both feminine and masculine, used for the three languages in the corpora generation. Regarding the professions, Table 23 contains the professions used for English in this project, and the previous ones (Bartl et al., 2020) and (Azpillaga Rivera, 2021). Table 24 shows the lists of professions for Basque and Spanish used in (Azpillaga Rivera, 2021) and that have been modified in the present work (the modified ones were reviewed in Section 4.1 and can be found in Table 4).

Feminine subjects		
English	**Basque**	**Spanish**
She	Ane	Ella
This woman	Emakume hau	Esta mujer
My sister	Nire arreba	Mi hermana
My daughter	Nire alaba	Mi hija
My wife	Nire emaztea	Mi mujer
My girlfriend	Nire emaztegia	Mi novia
My mother	Nire ama	Mi madre
My aunt	Nire izeba	Mi tía
My mom	Nire amatxo	Mi mamá

Masculine subjects		
English	**Basque**	**Spanish**
He	Jon	El
This man	Gizon hau	Este hombre
My brother	Nire anaia	Mi hermano
My son	Nire semea	Mi hijo
My husband	Nire senarra	Mi marido
My boyfriend	Nire senargaia	Mi novio
My father	Nire aita	Mi padre
My uncle	Nire osaba	Mi tío
My dad	Nire aitatxo	Mi papá

Table 22: Relation of subjects used to generate the BEC-Pro sentences for English, Basque and Spanish.

Male professions	female %	Female professions	female %	Balanced professions	female %
taper	0,7	kindergarten teacher	98,7	salesperson	48,5
steelworker	0,9	dental hygienist	96,0	director of religious activities	48,6
mobile equipment mechanic	1,3	speech-language pathologist	95,8	crossingguard	48,6
bus mechanic	1,5	dental assistant	94,9	photographer	49,3
service technician	1,5	childcareworker	93,4	lifeguard	49,4
heating mechanic	1,5	medical records technician	93,3	lodging manager	49,5
electrical installer	1,6	secretary	93,2	healthcare practitioner	49,5
operating engineer	1,7	medical assistant	92,7	sales agent	49,7
loggingworker	1,8	hairdresser	92,3	mail clerk	49,8
floor installer	1,9	dietitian	92,1	electrical assembler	50,4
roofer	1,9	vocational nurse	90,8	insurance sales agent	50,6
mining machine operator	2,0	teacher assistant	89,7	insurance underwriter	51,1
electrician	2,2	paralegal	89,6	medical scientist	51,8
repairer	2,2	billing clerk	89,5	statistician	52,5
conductor	2,4	phlebotomist	89,3	training specialist	52,5
plumber	2,7	receptionist	89,3	judge	52,5
carpenter	2,8	housekeeper	89,0	bartender	53,1
security system installer	2,9	registered nurse	88,9	dispatcher	53,1
mason	3,0	bookkeeper	88,5	order clerk	53,3
firefighter	3,3	health aide	83,3	mail sorter	53,3

Table 23: List of professions used for English.

Male professions		Female professions		Balanced professions	
suhiltzailea	bombero/a	haurtzaindegiko irakaslea	profesor/a de guardería	posta sailkatzailea	clasificador/a de correo
igeltseroa	albañil	hortzetako higienista	higienista dental	eskaera kudeatzailea	encargado/a de pedidos
segurtasun sistemako instalatzailea	instalador/a de sistemas de seguridad	logopeda	logopeda	operadorea	operador/a
arotza	carpintero/a	dentistaren laguntzailea	asistente dental	tabernaria	camarero/a
iturgina	fontanero/a	haur-zaintzailea	cuidador/a infantil	epailea	juez/a
tren-gidaria	maquinista	osasun txostenen teknikaria	técnico/a de expedientes médicos	entrenatzailea	entrenador
karrozaegilea	carrocero/a	idazkaria	secretario/a	estatistikaria	estadístico/a
elektrizista	electricista	mediku laguntzailea	asistente médico/a	mediku zientzialaria	científico/a médico/a
meatzaritzako makina-operadorea	operador/a de maquinaria minera	ile-apaintzailea	peluquero/a	aseguru-bitartekaria	asegurador/a
teilatu-emailea	techador/a	dietista	dietista	aseguruen salmenta agentea	agente de seguros
zoru-instalatzailea	instalador/a de suelos	erizain-laguntzailea	auxiliar de enfermería	muntatzaile-elektrikoa	montador/a eléctrico/a
egurketaria	trabajador/a maderera	irakasle laguntzailea	maestro/a ayudante	postako enplegatua	empleado/a de correos
eraikuntza ekipoen operadorea	operador/a de equipos de construcción	laguntzaile juridikoa	asistente legal	salmenta agentea	agente de ventas
instalatzaile-elektrikoa	instalador/a eléctrica	fakturazio-langilea	empleado/a de facturación	osasun-langilea	profesional sanitario/a
berogailu -teknikaria	técnico/a de calefacción	flebotomista	flebotomista	ostatu-arduraduna	encargado/a de alojamiento
auto mekanikaria	mecánico/a de automóviles	harreragilea	recepcionista	soroslea	salvavidas
autobus mekanikaria	mecánico/a autobuses	garbitzailea	limpiador/a	argazkilaria	fotógrafo/a
ekipamendu mugikorreko mekanikaria	mecánico/a de equipos móviles	erizain titularra	enfermero/a diplomado/a	pasabide-zaindaria	guardia de tráfico
altzairu-langilea	trabajador/a del acero	kontularia	contable	erlijio-jardueren zuzendaria	director/a de actividades religiosas
igeltsuzko panel instalatzailea	instalador/a de paneles de yeso	osasun laguntzailea	asistente sanitario/a	saltzailea	vendedor/a

Table 24: List of professions for Basque and Spanish used in the work of (Azpillaga Rivera, 2021). Note that this table does not contain percentages of female presence because the professions are the result of the English list translation.

B Tables of results

This appendix contains the complete tables of results for the three languages. There is one table for each language but, due to their length, they take 3 pages each. For each profession, the tables show in a more elaborated way the count, the association and the association removing the logarithm from the mathematical formula; everything for both types of subjects (F, M).

	BERT base count percentages					BERT base Association with Log						BERT base Association without Log					
Profession	**PosF %**	**PosM %**	**NegF %**	**NegM %**	**PM-PF %**	**PosF**	**PosM**	**NegF**	**NegM**	**AsocF**	**AsocM**	**PosF**	**PosM**	**NegF**	**NegM**	**AsocF**	**AsocM**
taper (M)	4,44	4,44	45,56	45,56	0,00	0,157	0,315	-1,241	-1,064	-1,117	-0,941	1,173	1,373	0,387	0,468	0,457	0,548
steelworker	10,00	30,00	40,00	20,00	20,00	1,479	0,913	-1,101	-0,652	-0,585	0,287	20,06	7,988	0,4	0,595	4,332	5,031
mobile equipment mechanic	7,78	32,22	42,22	17,78	24,44	1,874	0,976	-0,918	-0,588	-0,484	0,420	12,585	6,939	0,46	0,615	2,346	4,690
bus mechanic	5,56	27,78	44,44	22,22	22,22	1,255	0,982	-1,667	-0,691	-1,342	0,238	8,256	5,433	0,287	0,56	1,172	3,267
service technician	20,00	28,89	30,00	21,11	8,89	0,862	0,739	-0,514	-0,364	0,036	0,273	7,201	3,13	0,631	0,726	3,259	2,115
heating mechanic	6,67	25,56	43,33	24,44	18,89	1,086	0,927	-1,641	-0,826	-1,277	0,070	7,179	5,223	0,293	0,525	1,211	2,926
electrical installer	10,00	21,11	40,00	28,89	11,11	0,321	0,612	-1	-0,695	-0,736	-0,143	1,448	2,33	0,471	0,636	0,666	1,351
operating engineer	8,89	34,44	41,11	15,56	25,56	1,153	0,898	-1,209	-0,585	-0,789	0,437	14,491	5,826	0,399	0,608	2,904	4,203
logging worker	10,00	33,33	40,00	16,67	23,33	1,269	0,908	-0,954	-0,369	-0,509	0,482	18,551	7,167	0,462	0,718	4,080	5,017
floor installer	12,22	18,89	37,78	31,11	6,67	0,366	0,556	-1,008	-0,72	-0,672	-0,238	1,638	2,268	0,467	0,608	0,753	1,235
roofer	13,33	20,00	36,67	30,00	6,67	0,565	0,735	-0,838	-0,601	-0,464	-0,067	1,906	2,512	0,521	0,646	0,890	1,392
mining machine operator	7,78	34,44	42,22	15,56	26,67	1,44	1,002	-1,283	-0,909	-0,859	0,407	9,151	8,631	0,345	0,496	1,715	6,100
electrician	5,56	26,67	44,44	23,33	21,11	0,874	0,843	-0,992	-0,666	-0,785	0,139	3,704	3,507	0,424	0,57	0,788	2,136
repairer	5,56	27,78	44,44	22,22	22,22	0,514	0,872	-1,275	-0,586	-1,076	0,224	2,392	3,36	0,391	0,617	0,613	2,141
conductor	3,33	34,44	46,67	15,56	31,11	0,177	0,838	-1,516	-0,871	-1,403	0,306	1,195	2,968	0,349	0,582	0,405	2,226
plumber	13,33	22,22	36,67	27,78	8,89	0,572	0,877	-0,913	-0,434	-0,517	0,149	2,44	3,173	0,474	0,683	0,998	1,790
carpenter	3,33	27,78	46,67	22,22	24,44	1,719	0,909	-1,816	-0,58	-1,580	0,247	7,748	3,93	0,264	0,62	0,763	2,459
security system installer	11,11	27,78	38,89	22,22	16,67	0,628	0,691	-0,962	-0,786	-0,609	0,035	2,348	3,11	0,525	0,619	0,930	2,003
mason	6,67	33,33	43,33	16,67	26,67	1,44	0,861	-1,068	-0,392	-0,734	0,443	8,889	3,685	0,41	0,697	1,541	2,689
firefighter	10,00	31,11	40,00	18,89	21,11	0,896	0,827	-1,659	-0,488	-1,148	0,330	4,125	3,195	0,29	0,649	1,057	2,233

	BERT base count percentages					BERT base Association with Log						BERT base Association without Log					
kindergarten teacher (F)	36,67	6,67	13,33	43,33	-30,00	1,199	1,013	-0,364	-1,193	0,782	-0,899	25,783	4,249	0,718	0,401	19,099	0,914
dental hygienist	20,00	21,11	30,00	28,89	1,11	0,996	0,76	-0,586	-0,825	0,047	-0,156	9,789	3,404	0,661	0,546	4,312	1,753
speech-language pathologist	27,78	32,22	22,22	17,78	4,44	0,885	0,699	-0,799	-1,11	0,137	0,056	5,35	3,506	0,532	0,397	3,209	2,401
dental assistant	41,11	20,00	8,89	30,00	-21,11	0,965	0,716	-0,239	-0,61	0,751	-0,080	10,969	5,475	0,808	0,594	9,163	2,546
childcare worker	42,22	6,67	7,78	43,33	-35,56	1,017	0,882	-0,453	-1,076	0,788	-0,815	14,945	4,649	0,752	0,4	12,737	0,967
medical records technician	23,33	24,44	26,67	25,56	1,11	1,091	0,97	-0,515	-0,539	0,234	0,199	15,901	5,609	0,66	0,632	7,772	3,065
secretary	47,78	13,33	2,22	36,67	-34,44	0,963	0,887	-0,398	-1,196	0,903	-0,641	5,315	3,268	0,673	0,422	5,109	1,181
medical assistant	38,89	18,89	11,11	31,11	-20,00	0,998	0,94	-0,25	-0,547	0,721	0,015	11,807	6,029	0,795	0,634	9,360	2,672
hairdresser	34,44	12,22	15,56	37,78	-22,22	1,118	0,806	-0,871	-1,15	0,499	-0,672	6,786	2,982	0,503	0,399	4,831	1,030
dietitian	18,89	26,67	31,11	23,33	7,78	0,72	0,845	-1,12	-0,993	-0,425	-0,013	4,345	3,976	0,422	0,472	1,904	2,341
vocational nurse	47,78	3,33	2,22	46,67	-44,44	1,272	1,018	-0,264	-2,191	1,204	-1,977	22,82	3,213	0,768	0,218	21,840	0,418
teacher assistant	43,33	11,11	6,67	38,89	-32,22	1,044	0,801	-0,17	-0,979	0,882	-0,583	14,378	4,201	0,852	0,47	12,575	1,299
paralegal	21,11	24,44	28,89	25,56	3,33	0,622	0,624	-0,533	-0,684	-0,045	-0,045	2,193	2,115	0,671	0,589	1,314	1,335
billing clerk	21,11	18,89	28,89	31,11	-2,22	0,899	1,016	-0,452	-0,527	0,118	0,056	10,338	6,362	0,68	0,651	4,758	2,808
phlebotomist	18,89	30,00	31,11	20,00	11,11	0,578	0,833	-1,186	-1,45	-0,520	-0,080	2,385	3,287	0,434	0,392	1,171	2,129
receptionist	50,00	2,22	0,00	47,78	-47,78	1,243	0,72	0	-2,194	1,243	-2,064	8,304	2,367	0	0,188	8,304	0,285
housekeeper	43,33	1,11	6,67	48,89	-42,22	1,469	1,182	-0,779	-3,551	1,169	-3,446	16,699	3,26	0,509	0,058	14,540	0,129
registered nurse	45,56	3,33	4,44	46,67	-42,22	1,255	0,956	-0,618	-2,613	1,089	-2,375	30,203	2,991	0,567	0,145	27,569	0,335
bookkeeper	16,67	28,89	33,33	21,11	12,22	0,78	0,804	-1,03	-0,568	-0,427	0,225	5,027	3,37	0,484	0,636	1,998	2,216
health aide	44,44	13,33	5,56	36,67	-31,11	0,91	0,769	-0,284	-0,771	0,777	-0,360	9,904	3,714	0,764	0,509	8,888	1,364

	BERT base count percentages					BERT base Association with Log						BERT base Association without Log					
salesperson (B)	28,89	18,89	21,11	31,11	-10,00	0,705	0,845	-0,528	-0,643	0,184	-0,081	3,373	3,026	0,632	0,593	2,216	1,512
director of religious activities	17,78	22,22	32,22	27,78	4,44	1,512	1,106	-1,254	-1,541	-0,271	-0,365	20,596	6,268	0,386	0,329	7,572	2,969
crossing guard	6,67	35,56	43,33	14,44	28,89	0,443	0,794	-1,106	-0,324	-0,899	0,471	1,637	2,609	0,42	0,732	0,582	2,067
photographer	16,67	28,89	33,33	21,11	12,22	0,958	0,855	-0,972	-0,578	-0,329	0,250	5,789	3,782	0,449	0,589	2,229	2,434
lifeguard	15,56	26,67	34,44	23,33	11,11	0,966	0,927	-0,797	-0,395	-0,249	0,310	4,702	3,385	0,535	0,717	1,831	2,140
lodging manager	20,00	32,22	30,00	17,78	12,22	0,991	0,892	-0,857	-0,445	-0,118	0,417	18,524	5,267	0,546	0,688	7,737	3,639
healthcare practitioner	22,22	22,22	27,78	27,78	0,00	0,904	0,732	-0,739	-0,833	-0,009	-0,137	10,646	3,109	0,573	0,55	5,050	1,687
sales agent	22,22	15,56	27,78	34,44	-6,67	0,961	1,048	-0,556	-0,648	0,118	-0,120	11,174	7,141	0,62	0,589	5,311	2,627
mail clerk	22,22	21,11	27,78	28,89	-1,11	0,93	1,002	-0,602	-0,562	0,079	0,098	14,318	6,373	0,604	0,635	6,699	3,058
electrical assembler	2,22	23,33	47,78	26,67	21,11	0,342	0,458	-1,171	-0,929	-1,104	-0,282	1,408	1,895	0,435	0,594	0,478	1,201
insurance sales agent	16,67	20,00	33,33	30,00	3,33	1,337	1,239	-1,193	-0,884	-0,350	-0,035	19,411	10,582	0,409	0,526	6,743	4,548
insurance underwriter	12,22	23,33	37,78	26,67	11,11	1,122	0,85	-1,085	-1,003	-0,546	-0,138	16,589	4,724	0,44	0,523	4,388	2,483
medical scientist	12,22	32,22	37,78	17,78	20,00	1,053	0,832	-0,877	-0,506	-0,405	0,356	10,68	4,579	0,473	0,648	2,968	3,181
statistician	7,78	24,44	42,22	25,56	16,67	0,593	0,755	-1,211	-1,103	-0,930	-0,195	2,3	2,682	0,429	0,502	0,720	1,568
training specialist	14,44	25,56	35,56	24,44	11,11	0,944	0,801	-0,81	-0,484	-0,303	0,173	11,442	3,497	0,532	0,654	3,684	2,107
judge	6,67	34,44	43,33	15,56	27,78	1,145	0,709	-1,311	-0,884	-0,984	0,213	6,367	2,825	0,399	0,533	1,195	2,112
bartender	16,67	25,56	33,33	24,44	8,89	0,759	0,824	-0,828	-0,57	-0,299	0,142	3,886	3,243	0,501	0,608	1,629	1,955
dispatcher	12,22	33,33	37,78	16,67	21,11	0,632	0,673	-0,67	-0,472	-0,352	0,291	2,606	2,461	0,579	0,658	1,074	1,860
orderclerk	24,44	20,00	25,56	30,00	-4,44	0,921	0,89	-0,472	-0,538	0,209	0,033	12,907	4,254	0,653	0,639	6,644	2,085
mailsorter	13,33	26,67	36,67	23,33	13,33	0,478	0,453	-0,789	-0,69	-0,451	-0,080	1,857	1,908	0,567	0,606	0,911	1,300

Table 25: Full list of results obtained for English

	BERTeus count percentages					BERTeus Association with Log						BERTeus Association without Log					
Profession	PosF %	PosM %	NegF %	NegM %	PM-PF %	PosF	PosM	NegF	NegM	AsocF	AsocM	PosF	PosM	NegF	NegM	AsocF	AsocM
zuzendari (M)	20,00	24,44	30,00	25,56	4,44	0,53	0,55	-0,65	-0,56	-0,176	-0,019	1,897	1,912	0,599	0,651	1,12	1,27
kale-garbitzaile	28,89	23,33	21,11	26,67	-5,56	0,71	0,78	-0,62	-0,62	0,149	0,032	2,385	2,569	0,598	0,616	1,63	1,53
banatzaile	28,89	33,33	21,11	16,67	4,44	0,541	0,548	-0,435	-0,502	0,129	0,198	1,947	2,223	0,71	0,646	1,42	1,70
gidari	27,78	34,44	22,22	15,56	6,67	0,638	0,676	-0,429	-0,616	0,164	0,274	2,372	2,321	0,693	0,605	1,63	1,79
ingeniari	25,56	27,78	24,44	22,22	2,22	0,535	0,581	-0,603	-0,413	-0,021	0,139	1,996	2,10	0,597	0,697	1,31	1,48
polizi	20,00	24,44	30,00	25,56	4,44	0,487	0,439	-0,602	-0,443	-0,166	-0,012	1,777	1,758	0,595	0,687	1,07	1,21
programatzaile	28,89	30,00	21,11	20,00	1,11	0,519	0,593	-0,615	-0,508	0,040	0,153	1,903	2,243	0,595	0,652	1,35	1,61
lorezain	32,22	23,33	17,78	26,67	-8,89	0,616	0,547	-0,551	-0,541	0,201	-0,033	2,181	2,164	0,607	0,634	1,62	1,35
abeltzain	22,22	30,00	27,78	20,00	7,78	0,454	0,531	-0,561	-0,403	-0,110	0,157	1,674	1,983	0,611	0,714	1,08	1,48
peoi	20,00	22,22	30,00	27,78	2,22	0,465	0,571	-0,511	-0,377	-0,121	0,044	1,834	2,714	0,648	0,727	1,12	1,61
pilotu	34,44	38,89	15,56	11,11	4,44	0,537	0,598	-0,575	-0,609	0,191	0,330	2,116	2,13	0,606	0,621	1,65	1,79
meatzari	22,22	31,11	27,78	18,89	8,89	0,492	0,622	-0,452	-0,433	-0,032	0,223	1,774	2,252	0,69	0,681	1,17	1,66
suhiltzaile	25,56	30,00	24,44	20,00	4,44	0,475	0,526	-0,448	-0,429	0,024	0,144	1,765	2,018	0,684	0,695	1,24	1,49
beiragile	26,67	27,78	23,33	22,22	1,11	0,844	0,555	-0,452	-0,594	0,239	0,044	2,785	2,017	0,687	0,602	1,81	1,39
kamioilari	27,78	35,56	22,22	14,44	7,78	0,521	0,572	-0,49	-0,354	0,072	0,304	1,846	2,10	0,662	0,741	1,32	1,71
elektrikari	18,89	13,33	31,11	36,67	-5,56	0,537	0,715	-0,64	-0,831	-0,195	-0,419	1,852	2,237	0,57	0,521	1,05	0,98
mekanikari	26,67	26,67	23,33	23,33	0,00	0,62	0,55	-0,512	-0,472	0,092	0,073	2,027	2,154	0,643	0,663	1,38	1,46
arotz	20,00	23,33	30,00	26,67	3,33	0,444	0,567	-0,519	-0,505	-0,134	-0,005	1,679	2,374	0,641	0,635	1,06	1,45
iturgin	21,11	17,78	28,89	32,22	-3,33	0,592	0,573	-0,601	-0,65	-0,097	-0,215	2,061	2,246	0,602	0,583	1,22	1,17
igeltsero	20,00	24,44	30,00	25,56	4,44	0,42	0,488	-0,507	-0,423	-0,136	0,022	1,667	1,929	0,641	0,687	1,05	1,29

	BERTeus count percentages					BERTeus Association with Log						BERTeus Association without Log					
naturopata (F)	23,33	21,11	26,67	28,89	-2,22	0,621	0,812	-0,791	-0,687	-0,132	-0,054	2,355	2,959	0,509	0,555	1,37	1,57
etxe-langile	16,67	23,33	33,33	26,67	6,67	0,679	0,581	-0,588	-0,653	-0,166	-0,077	2,217	2,397	0,621	0,584	1,15	1,43
umezain	25,56	22,22	24,44	27,78	-3,33	0,567	0,553	-0,566	-0,481	0,013	-0,021	1,954	2,291	0,604	0,671	1,29	1,39
estetizista	21,11	11,11	28,89	38,89	-10,00	0,495	0,80	-0,946	-0,988	-0,338	-0,591	1,786	2,804	0,463	0,455	1,02	0,98
ile-apaintzaile	24,44	21,11	25,56	28,89	-3,33	0,606	0,584	-0,555	-0,781	0,013	-0,205	2,028	2,119	0,623	0,517	1,31	1,19
garbitzaile	16,67	24,44	33,33	25,56	7,78	0,527	0,444	-0,588	-0,581	-0,216	-0,080	1,951	1,885	0,602	0,599	1,05	1,23
erizain	17,78	27,78	32,22	22,22	10,00	0,492	0,638	-0,688	-0,48	-0,268	0,141	1,811	2,232	0,554	0,662	1,00	1,53
teleoperadore	15,56	21,11	34,44	28,89	5,56	0,608	0,567	-0,63	-0,798	-0,245	-0,222	1,976	2,084	0,585	0,534	1,02	1,19
kutxazain	12,22	15,56	37,78	34,44	3,33	0,722	0,70	-0,783	-0,587	-0,415	-0,187	2,307	2,906	0,525	0,616	0,96	1,33
txartel-saltzaile	30,00	27,78	20,00	22,22	-2,22	0,596	0,668	-0,714	-0,833	0,072	0,001	2,206	2,251	0,526	0,508	1,53	1,48
psikologo	18,89	31,11	31,11	18,89	12,22	0,424	0,544	-0,653	-0,509	-0,246	0,146	1,714	1,917	0,577	0,663	1,01	1,44
historialari	15,56	28,89	34,44	21,11	13,33	0,564	0,425	-0,503	-0,464	-0,171	0,050	2,206	1,715	0,661	0,674	1,14	1,28
gizarte-langile	14,44	21,11	35,56	28,89	6,67	0,756	0,642	-0,649	-0,676	-0,243	-0,120	2,363	3,06	0,582	0,567	1,10	1,62
liburuzain	11,11	18,89	38,89	31,11	7,78	0,517	0,604	-0,681	-0,546	-0,415	-0,112	1,899	2,075	0,565	0,619	0,86	1,17
farmazialari	23,33	23,33	26,67	26,67	0,00	0,73	0,668	-0,519	-0,478	0,064	0,057	2,273	2,333	0,647	0,669	1,41	1,45
saltzaile	25,56	28,89	24,44	21,11	3,33	0,537	0,503	-0,375	-0,358	0,091	0,139	1,939	2,096	0,739	0,733	1,35	1,52
odontologo	25,56	25,56	24,44	24,44	0,00	0,551	0,68	-0,696	-0,992	-0,059	-0,137	2,018	2,284	0,573	0,452	1,31	1,39
harreragile	30,00	28,89	20,00	21,11	-1,11	0,57	0,526	-0,614	-0,457	0,096	0,111	2,051	1,992	0,592	0,684	1,47	1,44
idazkari	20,00	33,33	30,00	16,67	13,33	0,565	0,554	-0,546	-0,588	-0,102	0,173	1,921	1,948	0,638	0,631	1,15	1,51
kazetari	14,44	24,44	35,56	25,56	10,00	0,605	0,478	-0,478	-0,414	-0,165	0,022	2,072	1,882	0,679	0,704	1,08	1,28

	BERTeus count percentages					BERTeus Association with Log						BERTeus Association without Log					
mediku (B)	16,67	28,89	33,33	21,11	12,22	0,553	0,527	-0,549	-0,512	-0,182	0,088	1,956	1,897	0,636	0,641	1,08	1,37
epaile	17,78	25,56	32,22	24,44	7,78	0,467	0,433	-0,682	-0,472	-0,273	-0,009	1,755	1,66	0,558	0,673	0,98	1,18
biologo	21,11	27,78	28,89	22,22	6,67	0,54	0,655	-0,532	-0,471	-0,081	0,155	1,916	2,287	0,65	0,671	1,18	1,57
albaitari	22,22	26,67	27,78	23,33	4,44	0,527	0,524	-0,613	-0,342	-0,106	0,120	1,905	1,827	0,606	0,742	1,18	1,32
sukaldari	22,22	33,33	27,78	16,67	11,11	0,469	0,46	-0,484	-0,411	-0,060	0,170	1,739	1,797	0,658	0,715	1,14	1,44
zerbitzari	20,00	26,67	30,00	23,33	6,67	0,563	0,547	-0,37	-0,397	0,003	0,106	1,964	2,324	0,73	0,694	1,22	1,56
postari	24,44	22,22	25,56	27,78	-2,22	0,589	0,629	-0,528	-0,64	0,018	-0,076	1,928	2,289	0,646	0,572	1,27	1,34
higiezin-agente	31,11	22,22	18,89	27,78	-8,89	0,615	0,569	-0,753	-0,647	0,098	-0,107	2,133	1,998	0,545	0,583	1,53	1,21
kontulari	21,11	22,22	28,89	27,78	1,11	0,455	0,551	-0,718	-0,622	-0,223	-0,101	1,685	2,166	0,542	0,578	1,02	1,28
abokatu	16,67	27,78	33,33	22,22	11,11	0,506	0,491	-0,58	-0,465	-0,218	0,066	1,882	1,791	0,612	0,669	1,04	1,29
hornitzaile	26,67	33,33	23,33	16,67	6,67	0,588	0,587	-0,489	-0,454	0,085	0,240	2,008	2,317	0,68	0,691	1,39	1,78
mediku-bisitari	13,33	24,44	36,67	25,56	11,11	0,706	0,63	-0,681	-0,898	-0,311	-0,151	2,228	2,17	0,565	0,477	1,01	1,30
argazkilari	21,11	25,56	28,89	24,44	4,44	0,468	0,549	-0,523	-0,481	-0,105	0,045	1,774	1,983	0,653	0,671	1,13	1,34
dekoratzaile	24,44	20,00	25,56	30,00	-4,44	0,649	0,567	-0,659	-0,80	-0,020	-0,253	2,233	2,047	0,595	0,534	1,40	1,14
entrenatzaile	16,67	27,78	33,33	22,22	11,11	0,57	0,511	-0,612	-0,616	-0,218	0,010	2,056	1,869	0,587	0,60	1,08	1,31
gozogile	18,89	28,89	31,11	21,11	10,00	0,641	0,49	-0,523	-0,478	-0,083	0,081	2,188	1,941	0,642	0,661	1,23	1,40
fisikari	25,56	27,78	24,44	22,22	2,22	0,483	0,554	-0,632	-0,438	-0,062	0,113	1,844	1,98	0,584	0,692	1,23	1,41
matematikari	20,00	22,22	30,00	27,78	2,22	0,488	0,703	-0,626	-0,492	-0,180	0,039	1,92	2,41	0,587	0,668	1,12	1,44
inkestatzaile	26,67	30,00	23,33	20,00	3,33	0,77	0,672	-0,562	-0,46	0,148	0,219	2,396	2,343	0,639	0,673	1,58	1,68
bedel	20,00	23,33	30,00	26,67	3,33	0,662	0,38	-0,643	-0,75	-0,121	-0,223	2,134	1,55	0,592	0,535	1,21	1,01

Table 26: Full list of results obtained for Basque

	BETO count percentages					BETO Association with Log						BETO Association without Log					
Profession	**PosF %**	**PosM %**	**NegF %**	**NegM %**	**PM-PF %**	**PosF**	**PosM**	**NegF**	**NegM**	**AsocF**	**AsocM**	**PosF**	**PosM**	**NegF**	**NegM**	**AsocF**	**AsocM**
directora (M)	37,78	27,78	12,22	22,22	-10,00	1,731	1,877	-0,628	-0,694	1,154	0,734	47,66	52,20	0,60	0,56	36,16	29,25
barrendera	41,11	33,33	8,89	16,67	-7,78	1,565	1,718	-0,480	-0,575	1,201	0,954	24,82	49,77	0,66	0,60	20,52	33,38
repartidora	45,56	38,89	4,44	11,11	-6,67	1,927	1,435	-0,214	-0,700	1,737	0,961	55,14	42,14	0,82	0,53	50,31	32,90
conductora	40,00	31,11	10,00	18,89	-8,89	2,056	1,863	-0,206	-0,878	1,604	0,828	62,46	56,76	0,83	0,45	50,13	35,48
ingeniera	45,56	40,00	4,44	10,00	-5,56	1,844	1,567	-0,411	-0,554	1,644	1,143	56,53	46,35	0,68	0,63	51,56	37,20
policía	30,00	31,11	20,00	18,89	1,11	2,023	1,846	-0,992	-0,537	0,817	0,946	50,82	41,25	0,49	0,64	30,69	25,91
programadora	44,44	34,44	5,56	15,56	-10,00	1,980	1,671	-0,326	-0,501	1,724	0,995	62,87	50,63	0,73	0,64	55,96	35,08
jardinera	38,89	43,33	11,11	6,67	4,44	2,168	1,504	-0,517	-0,577	1,571	1,227	59,42	43,55	0,63	0,63	46,36	37,83
ganadera	38,89	34,44	11,11	15,56	-4,44	1,573	1,736	-1,641	-0,978	0,859	0,892	17,06	38,89	0,41	0,48	13,36	26,94
peona	42,22	30,00	7,78	20,00	-12,22	1,774	1,717	-0,144	-0,723	1,476	0,741	34,66	41,46	0,88	0,53	29,41	25,09
piloto	24,44	32,22	25,56	17,78	7,78	1,760	1,675	-0,845	-0,404	0,429	0,936	16,55	48,36	0,51	0,71	8,35	31,42
minera	45,56	40,00	4,44	10,00	-5,56	1,448	1,664	-0,147	-0,561	1,306	1,219	22,99	49,73	0,87	0,65	21,02	39,91
bombera	47,78	38,89	2,22	11,11	-8,89	1,877	1,582	-0,212	-0,458	1,784	1,129	48,42	53,08	0,81	0,69	46,31	41,44
cristalera	46,67	38,89	3,33	11,11	-7,78	1,623	1,537	-0,257	-0,359	1,498	1,116	22,24	36,78	0,78	0,73	20,80	28,77
camionera	46,67	40,00	3,33	10,00	-6,67	2,030	1,660	-0,291	-0,632	1,875	1,202	64,54	53,94	0,75	0,59	60,29	43,27
electricista	22,22	34,44	27,78	15,56	12,22	2,297	1,731	-0,633	-0,557	0,669	1,019	49,65	50,66	0,57	0,62	22,38	35,09
mecánica	41,11	37,78	8,89	12,22	-3,33	1,773	1,697	-0,357	-0,550	1,394	1,148	42,41	54,03	0,71	0,62	34,99	40,97
carpintera	46,67	43,33	3,33	6,67	-3,33	2,002	1,588	-0,050	-0,598	1,865	1,297	72,02	47,06	0,95	0,59	67,28	40,86
fontanera	44,44	33,33	5,56	16,67	-11,11	2,037	1,879	-0,377	-0,584	1,769	1,058	64,35	60,74	0,71	0,61	57,28	40,69
albañil	30,00	34,44	20,00	15,56	4,44	1,751	1,677	-0,648	-0,617	0,791	0,963	37,87	45,48	0,59	0,61	22,96	31,52

	BETO count percentages					BETO Association with Log						BETO Association without Log					
naturópata (F)	38,89	32,22	11,11	17,78	-6,67	1,668	1,694	-0,412	-0,657	1,206	0,858	31,23	38,40	0,70	0,56	24,45	24,95
empleada doméstica	46,67	32,22	3,33	17,78	-14,44	1,901	1,844	-0,111	-0,972	1,767	0,843	47,77	44,43	0,90	0,44	44,64	28,79
niñera	43,33	36,67	6,67	13,33	-6,67	1,714	1,693	-0,365	-0,580	1,437	1,087	21,45	46,35	0,71	0,60	18,69	34,15
esteticista	32,22	33,33	17,78	16,67	1,11	1,907	1,637	-0,545	-0,798	1,035	0,825	43,59	45,61	0,65	0,50	28,32	30,57
peluquera	48,89	42,22	1,11	7,78	-6,67	1,954	1,597	-0,027	-0,813	1,910	1,222	63,79	55,20	0,97	0,52	62,39	46,70
limpiadora	50,00	36,67	0,00	13,33	-13,33	1,740	1,582	0,000	-0,442	1,740	1,042	49,83	51,39	0,00	0,68	49,83	37,86
enfermera	44,44	40,00	5,56	10,00	-4,44	1,810	1,632	-0,447	-0,468	1,559	1,212	55,89	48,33	0,67	0,70	49,75	38,81
teleoperadora	44,44	35,56	5,56	14,44	-8,89	1,576	1,595	-0,261	-0,654	1,372	0,945	37,78	47,73	0,78	0,56	33,67	34,10
cajera	44,44	33,33	5,56	16,67	-11,11	1,968	1,729	-0,230	-0,513	1,724	0,982	75,64	58,83	0,81	0,64	67,32	39,43
taquillera	48,89	34,44	1,11	15,56	-14,44	1,659	1,654	-0,415	-0,556	1,613	0,966	43,74	52,67	0,66	0,62	42,78	36,48
psicóloga	45,56	37,78	4,44	12,22	-7,78	1,964	1,700	-0,587	-0,830	1,737	1,082	73,39	49,14	0,61	0,50	66,92	37,25
historiadora	44,44	35,56	5,56	14,44	-8,89	1,873	1,720	-1,299	-0,893	1,521	0,965	59,31	39,47	0,31	0,50	52,75	28,21
trabajadora social	41,11	32,22	8,89	17,78	-8,89	1,968	1,915	-0,549	-0,852	1,521	0,931	63,26	57,34	0,63	0,51	52,12	37,13
bibliotecaria	46,67	32,22	3,33	17,78	-14,44	1,931	1,774	-0,443	-0,520	1,773	0,958	57,57	47,43	0,69	0,65	53,78	30,80
farmacéutica	45,56	38,89	4,44	11,11	-6,67	1,607	1,617	-0,188	-0,653	1,447	1,113	37,44	48,10	0,83	0,55	34,18	37,53
dependienta	42,22	33,33	7,78	16,67	-8,89	1,833	1,279	-0,545	-0,741	1,463	0,606	45,35	13,01	0,65	0,56	38,39	8,86
dentista	30,00	34,44	20,00	15,56	4,44	1,427	1,792	-0,593	-0,358	0,619	1,123	12,83	53,66	0,60	0,74	7,94	37,19
recepcionista	34,44	21,11	15,56	28,89	-13,33	2,138	1,886	-0,425	-1,093	1,341	0,165	51,48	26,43	0,68	0,42	35,68	11,40
secretaria	38,89	30,00	11,11	20,00	-8,89	1,956	1,601	-0,478	-0,578	1,415	0,729	56,80	24,63	0,66	0,60	44,32	15,02
periodista	31,11	32,22	18,89	17,78	1,11	1,791	1,748	-0,732	-0,774	0,838	0,851	46,32	39,51	0,57	0,53	29,04	25,65

	BETO count percentages					BETO Association with Log						BETO Association without Log					
médica (B)	44,44	38,89	5,56	11,11	-5,56	1,785	1,671	-0,262	-0,626	1,558	1,161	42,56	53,27	0,78	0,60	37,92	41,56
jueza	41,11	31,11	8,89	18,89	-10,00	1,938	1,797	-0,801	-0,924	1,451	0,769	71,02	52,12	0,49	0,49	58,48	32,61
bióloga	44,44	36,67	5,56	13,33	-7,78	2,091	1,702	-0,623	-0,586	1,789	1,092	53,55	43,33	0,59	0,63	47,66	31,94
veterinaria	45,56	35,56	4,44	14,44	-10,00	1,984	1,726	-0,188	-0,543	1,791	1,071	47,56	55,95	0,84	0,64	43,41	39,97
cocinera	48,89	37,78	1,11	12,22	-11,11	1,963	1,679	-0,219	-0,483	1,915	1,151	58,01	47,36	0,80	0,66	56,74	35,94
camarera	43,33	33,33	6,67	16,67	-10,00	2,000	1,824	-0,210	-0,692	1,705	0,985	63,34	59,43	0,82	0,55	55,01	39,80
cartera	35,56	38,89	14,44	11,11	3,33	0,927	1,684	-0,320	-0,587	0,567	1,179	3,64	52,15	0,75	0,64	2,80	40,70
agente inmobiliaria	40,00	26,67	10,00	23,33	-13,33	1,793	1,678	-0,281	-1,133	1,378	0,366	47,03	28,30	0,77	0,47	37,77	15,31
contable	32,22	32,22	17,78	17,78	0,00	1,710	1,631	-0,692	-0,670	0,856	0,813	39,20	39,13	0,58	0,56	25,46	25,42
abogada	45,56	35,56	4,44	14,44	-10,00	1,810	1,714	-0,469	-0,757	1,607	1,000	74,40	54,16	0,67	0,55	67,84	38,67
reponedora	40,00	33,33	10,00	16,67	-6,67	1,731	1,471	-0,328	-0,860	1,319	0,694	28,32	22,95	0,74	0,51	22,81	15,47
visitadora médico	46,67	35,56	3,33	14,44	-11,11	1,785	1,632	-1,144	-0,962	1,590	0,883	36,18	26,56	0,33	0,43	33,79	19,01
fotógrafa	43,33	38,89	6,67	11,11	-4,44	2,054	1,768	-0,659	-0,879	1,692	1,180	72,73	51,65	0,57	0,51	63,11	40,28
decoradora	47,78	40,00	2,22	10,00	-7,78	1,992	1,718	-0,269	-0,715	1,892	1,231	59,82	50,47	0,77	0,53	57,19	40,48
entrenadora	40,00	26,67	10,00	23,33	-13,33	1,850	1,991	-0,467	-0,630	1,387	0,768	61,47	62,50	0,64	0,62	49,30	33,62
panadera	47,78	41,11	2,22	8,89	-6,67	2,016	1,681	-0,211	-0,522	1,917	1,289	59,08	47,69	0,82	0,65	56,49	39,33
física	41,11	40,00	8,89	10,00	-1,11	1,587	1,510	-0,355	-0,571	1,242	1,094	28,14	34,49	0,72	0,64	23,27	27,72
matemática	44,44	38,89	5,56	11,11	-5,56	1,520	1,585	-0,391	-0,642	1,308	1,090	38,43	43,19	0,73	0,59	34,24	33,72
encuestadora	43,33	33,33	6,67	16,67	-10,00	1,872	1,683	-0,439	-0,447	1,564	0,973	39,96	46,86	0,68	0,68	34,72	31,46
bedel	21,11	32,22	28,89	17,78	11,11	1,744	1,310	-0,792	-0,396	0,279	0,703	15,51	13,23	0,51	0,69	6,85	8,77

Table 27: Full list of results obtained for Spanish