eman ta zabal zazu

Universidad del País Vasco | Euskal Herriko Unibertsitatea

# Labour Statistics vs. Static Word Embeddings: A Comparison of Gender Bias

Author: Andrés Figueroa Vásquez

Advisors: Olatz Pérez de Viñaspre and Xabier Arregi

hap/lap

Hizkuntzaren Azterketa eta Prozesamendua

Language Analysis and Processing

Final Thesis

October 2021

## Laburpena

Proiektu honek lanbideen genero banaketa jasotzen duten inkesten informazioaren eta hiru hizkuntza-ereduen arteko erlazioa aztertzen du: GloVE, word2vec eta fastText, ingelesez zein gaztelaniaz. Helburua, mundu laboralaren eta hizkuntza ereduen arteko ezberdintasunak aztertzea izan da, genero-alborapenaren ikuspegitik. Horretarako, datu-multzo linguistiko batzuk sortu dira, beste lan batzuetako autoreek "muturreko emakume lanbideak" eta "muturreko gizonezko lanbideak" bezala izendatu dituztenak. Genero-alborapena hobeto aztertzeko asmoz, "lanbide neutroen" datu-multzoa ere sortu da konparaketa egiteko. Horrela, hizkuntza-eredu estatiko ezberdinak, teknika eta corpusen aldetik eta hizkuntza ezberdinak erabiltzeak dituen aldeak aztertu da, baita horien baitako patroiak ere.



**Hitz gakoak:** genero alborapena, hitz txertatze estatikoak, etika, adimen artifiziala



## Abstract

This project explores the relation between labour statistics information and three language models: GloVe, word2vec and fastText, in both English and Spanish. The aim is to see what differs in reality *versus* word embedding spaces in terms of gender bias. To do so, diverse linguistic data sets were created, using what previous authors called extreme *she* occupations and extreme *he* occupations. To better assess their behaviour, these outcomes were compared to *gender-neutral* professions. This way, the variation of utilising different static word embeddings, corpora and natural languages will be determined, as to discover the patterns that lie underneath them.



**Keywords:** gender bias, static word embeddings, ethics, artificial intelligence

# Contents

# List of Figures & Graphs

# List of Tables

# 1. Introduction

Word embeddings are widely used in NLP for a vast range of tasks. According to previous research, they are derived from text corpora and replicate the gender biases found in daily life.

This phenomenon is pervasive and consistent, provoking serious implications. The aim of this project is to analyse the gender bias of the word embeddings and language models, through the assessment of their effects in different corpora and languages.

The gender bias of a word $w$ is defined by some authors as its projection on the "gender direction":

$$w \cdot (he - she),$$

assuming all vectors are normalised. The larger a word's projection is on $he - she$, the more biased it is. They also quantify the bias in word embeddings using this definition and show it aligns well with social stereotypes. Some researchers think of it as a much deeper and systematic situation, and that simply reducing the projection of words on a gender direction is insufficient: it merely hides the bias.

On the other hand, the ability to automatically detect how the meaning of words evolves in this setting has potential value for research in Lexicography and Linguistics, but also in real life. This sort of knowledge would render word meaning representations more accurate, in terms of semantic information. This is a very important step, since it is desirable to have a method of assessment that is data driven, rather than based on intuitive judgments.

Word meanings vary according to the historical contexts where they develop, as well as the genre and register of the discourse. They exhibit a range of senses whose distribution or prevalence are related to these features. An automated procedure for extracting information from text would be useful for historical exploratory studies or information retrieval.

This thesis consists of exploring word embeddings and language models in English and Spanish to quantify the gender bias before applying de-biasing techniques.

In the following pages, there will be a theoretical description about word embeddings and gender bias (section 2). Then, the reader will find a methodology explanation for the experiments (section 3), along with a discussion about those findings (section 4), followed by the conclusions (section 5).

# 2. Literature review

## 2.1 Word Embeddings

The mathematical representation of words, sentences and other linguistics elements has been at the core of the Natural Language Processing (NLP) field right from its early start. To achieve its findings, vectors of real numbers have played a paramount role.

At first, they were used to summarise document contents for information retrieval (Salton et al., 1975). One important thing to note is that most of the research done on vector representation for NLP has had an emphasis on word vectors, also known as word embeddings (WE henceforth). In WE models, "each word in a given language is assigned to a high-dimensional vector such that the geometry of the vectors captures semantic relations between words" (Garg et al., 2018). For instance, vectors being closer together has been shown to correspond to more similar words. According to the *distributional hypothesis* (Harris, 1954), words that appear in the same contexts tend to have analogous senses. That is the base of most word embedding (WE) methods.

At this point, it becomes particularly significant to underline the prediction-based WE, which train by learning words that are likely to appear in a given context. Such a method was originally proposed by Bengio et al. (2003). Back then, WE were learnt through neural language modelling.

As a consequence, the universalisation of WE in NLP became a reality, thanks to the following prediction-based models: first *word2vec* (Mikolov et al., 2013a) and GloVe (Pennington et al., 2014), created a few years later.

Afterward, the contextualised language models were invented. They are one of the most important discoveries in NLP. Two of the most recognised sorts are ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018). These methods, also known as *contextualisers* (as Liu et al. call them), are examples of an effective application of *transfer learning*. Provided with a sequence of input words, they output a contextualised word embedding (CWE).

Whilst WE symbolise isolated word meanings, CWE characterise word meanings in context, by capturing some linguistic phenomena that work over the word level.

Although the extended adoption of WE was achieved, some obstacles needed to be overcome. Even before the beginning of the 21$^{st}$ century, Schütze (1998) had already pointed out that many words have not only one but several *meanings* or *senses*, that is to say, they are *polysemous*. Given the situation, it seems inaccurate to represent every word as a single vector in a semantic space, ignoring how many senses it could have.

In this regard, Camacho-Collados and Pilehvar (2018) coined the term "meaning conflation deficiency". In short, it distorts WE because of polysemy. For example, it would make unconnected words such as *left* and *wrong* arbitrarily near in the vector space because they are related in two different senses of the word *right*, as Neelakantan et al. (2014) found. Therefore, it would be expected to see this problem impeding the semantic understanding of WE models. Yaghoobzadeh & Schütze (2016) said that WE models can experience complications to determine which sense of an ambiguous term applies in certain contexts.

## 2.2 Language Models

### 2.2.1 Definition and Early Stages

A language model is a statistical representation that outputs a probability distribution over a vocabulary, given the beginning of a sequence of words (e.g. *"Why did the"*, if the full sequence is *"Why did the chicken cross the road?"*). It estimates the chances of each word in the vocabulary being the next to appear in the sentence structure. More formally, given words $w_1, w_2, \ldots, w_{t-1} \in V$, a language model produces a probability distribution over $V$ for $w_t$. Language modelling can be seen as a multiclass classification issue, where the samples are concatenations of words and the classes are the words in a vocabulary. Nonetheless, language modelling is thought to be an unsupervised task, since the samples can be built from an unlabelled text corpus.

At early stages, the prevailing methodology for language modelling was the n-gram approach, as described by Shannon (1951). An n-gram is a particular succession of $n$ words. The possibility foreseen by an n-gram model of a certain word appearing next in a series is the probability observed in the corpus of the word occurring, per the previous $N - 1$ tokens. A serious disadvantage of n-gram models is that they are not capable of generalising well to sequences that have not been seen beforehand. Rather, they assign higher odds only to the successions that have occurred in the training corpus. Since the number of probable n-grams is exponentially high, even a large corpus can only contain a small subset of them.

Later on, a competing methodology was crafted by Bengio et al. (2003). As a result, they offered the neural language model. Here *neural* refers to the use of continuous representations of word sequences rather than n-grams, that is to say, the use of artificial neural networks.

It consists of two components: an embedding matrix $E \in \mathbb{R}^{|V| \times d}$, where $d$ is the dimensionality of the embedding, and a probability function $f$. The neural model uses the embeddings of the previous words in the sequence as an input and outputs an estimation of the probability distribution over the next words.

If someone wants to find a fixed-sized input to $f$, this model only takes the $n$ previous words ($w_{t-n}$, $w_{t-n+1}$, ..., $w_{t-1}$) when calculating $w_t$. The resulted representation $x \in \mathbb{R}^{nd}$ is given by concatenating the embeddings of these words. For instance, $x = [e_{wt-n}; e_{wt-n+1}; ...; e_{wt-1}]$, where $e_i$ is the $i^{\text{th}}$ row of $E$.

This scheme reached a notorious performance at the time, by a noteworthy margin over the previous n-gram models. It was able to generalise in ways the other models could not. In this regard, Bengio et al. (2003) gave the following explanation:

> If we knew that *dog* and *cat* played similar roles (semantically and syntactically), and similarly for (*the,a*), (*bedroom,room*), (*is,was*), (*running,walking*), we could naturally generalize (i.e. transfer probability mass) from
>
> | | |
> |---|---|
> | | *The cat is walking in the bedroom* |
> | to | *A dog was running in a room* |
> | and likewise to | *The cat is running in a room* |
> | | *A dog is walking in a bedroom* |
> | | *The dog was walking in the room* |
> | | ... |
>
> and many other combinations. In the proposed model, it will so generalize because "similar" words are expected to have a similar feature vector, and because the probability function is a smooth function of these feature values, a small change in the features will induce a small change in the probability. Therefore, the presence of only one of the above sentences in the training data will increase the probability, not only of that sentence, but also of its combinatorial number of "neighbors" in sentence space (as represented by sequences of feature vectors).

Without a doubt, Bengio et al.'s work became a proof of all the capacity word embeddings (WE) had. However, it was only utilised as a way to obtain the best possible language model performance. Subsequent research introduced the notion that they could train on one specific task and, then, be eventually applied to another.

### 2.2.2 Static Models

In the following pages, there will be an explanation about the static models used in the course of this project, namely: fastText, GloVe and word2vec.

#### *fastText*

In words of its authors, this is "a simple and efficient baseline for text classification". It uses a bag of n-grams as additional feature to capture some partial information about the local word order (Joulin et al., 2016). It was evaluated on two different tasks. Firstly, there was a comparison with existing text classifiers at the time, related to sentiment analysis. Secondly, its capacity to scale to large output space on a tag prediction dataset was assessed. fastText obtained performance on par with proposed methods at that particular moment, inspired by deep learning, while being much faster (Joulin et al., 2016).

A year later, Bojanowski et al. (2017) offered a new approach based on the skip-gram model, where each word was represented as a bag of character n-grams. A vector representation was associated to each character n-gram; words being represented as the sum of these representations. Their method allowed to train models on large corpora quickly and was able to calculate word representations for words that did not appear in the training data. It evaluated their word representations on nine different languages, both on word similarity and analogy tasks. By comparing to morphological word representations, these vectors achieved state-of-the-art performance on these tasks.

Bojanowski et al. (2017) recommended their model to learn word representations whilst considering morphology. The morphology was modelled by considering sub-word units, and representing words by a sum of its character n-grams. They concluded that morphological information meaningfully developed the syntactic tasks, by outperforming the baselines.

In contrast, it did not help for semantic questions (especially analogies), and even degraded the performance for German and Italian. They argued that it was closely linked with the choice of the length of character n-grams that they utilised.

***Global Vector (GloVE)***

This is a model proposed by Pennington et al. (2014). Unlike Mikolov et al., 2013a, who mixed CBOW and the skip-gram model, they decided to combine the count-based matrix factorisation and the context-based skip-gram model together.

As explained before, the counts and co-occurrences of words can deduce their senses. In order to distinguish from $p(w_O|w_I)$ in the context of a WE, there has to be a co-occurrence probability:

$$p_{\text{co}}(w_k|w_i) = \frac{C(w_i, w_k)}{C(w_i)}$$

where $C(w_i,w_k)$ counts the co-occurrence between $w_i$ and $w_k$. So, given two words $w_i$ = "rock" and $w_j$ = "air", and a third word $\hat{w}_k$ = "solid" related to "rock" but not to "air", it is expected for $p_{co}(\hat{w}_k \mid w_i)$ to be much larger than $p_{co}(\hat{w}_k \mid w_j)$.

The intuition is that the meanings are captured by the ratios of co-occurrence probabilities rather than the probabilities themselves. GloVe models the connection between two terms regarding to the third context term as:

$$F(w_i, w_j, \tilde{w}_k) = \frac{p_{\text{co}}(\tilde{w}_k|w_i)}{p_{\text{co}}(\tilde{w}_k|w_j)}$$

Besides, since the objective is to learn meaningful word vectors, $F$ is meant to be a function of the linear difference between two words $w_i - w_j$:

$$F((w_i - w_j)^\top \tilde{w}_k) = \frac{p_{\text{co}}(\tilde{w}_k|w_i)}{p_{\text{co}}(\tilde{w}_k|w_j)}$$

F is symmetric between target words and context words, so the final solution is to model $F$ as an exponential function, as explained by Pennington et al. (2014).

$$F({w_i}^\top \tilde{w}_k) = \exp({w_i}^\top \tilde{w}_k) = p_{\text{co}}(\tilde{w}_k|w_i)$$
$$F((w_i - w_j)^\top \tilde{w}_k) = \exp((w_i - w_j)^\top \tilde{w}_k) = \frac{\exp(w_i^\top \tilde{w}_k)}{\exp(w_j^\top \tilde{w}_k)} = \frac{p_{\text{co}}(\tilde{w}_k|w_i)}{p_{\text{co}}(\tilde{w}_k|w_j)}$$

Finally,

$${w_i}^\top \tilde{w}_k = \log p_{\text{co}}(\tilde{w}_k|w_i) = \log \frac{C(w_i, \tilde{w}_k)}{C(w_i)} = \log C(w_i, \tilde{w}_k) - \log C(w_i)$$

Since the second term – log $C(w_i)$ is independent of $k$, a bias term $b_i$ for $w_i$ can be added in order to capture – log $C(w_i)$. To maintain a symmetric form, another bias $b_k$ can be added for $\hat{w}_k$.

$$\log C(w_i, \tilde{w}_k) = w_i{}^\top \tilde{w}_k + b_i + \tilde{b}_k$$

The loss function for the GloVe model is created to preserve the above-mentioned formula, by reducing the sum of the squared errors:

$$\mathcal{L}_\theta = \sum_{i=1,j=1}^{V} f(C(w_i, w_j))(w_i{}^\top \tilde{w}_j + b_i + \tilde{b}_j - \log C(w_i, \tilde{w}_j))^2$$

The weighting schema f(c) is a function of the co-occurrence of $w_i$ and $w_j$ and it is an adaptable configuration. It should be close to zero as $c \to 0$; should not decrease, since higher co-occurrence should have more impact and should saturate when $c$ become extremely large. The authors, then, suggested this weighting function:

$$f(c) = \begin{cases} (\frac{c}{c_{\max}})^\alpha & \text{if } c < c_{\max}, c_{\max} \text{ is adjustable.} \\ 1 & \text{if otherwise} \end{cases}$$

***word2vec***

A revolution in a more versatile kind of WE came through two combined methods, the continuous bag of words (CBOW) and skip-gram (Mikolov et al., 2013a), in the form of the *word2vec* software package. It mainly works based on simplicity and efficiency to empower quick training on large corpora.

The result is a neural network for classifying co-occurring words, which takes a term and its $d$ preceding and succeeding counterparts. Next, CBOW uses the neighbouring vocabulary and guesses the central word. At the same time, the skip-gram portion seizes the terminology and deductes its neighbourhood.

In both scenarios, the neural network shows similar ways of working: there was one word on the input (either the neighbouring term for CBOW or the central term for the skip-gram model) and one word on the output, embodied in one-hot encoding. Experimental assessments demonstrate a slim lead of the skip-gram model over CBOW for several tasks, that is why there is going to be an emphasis on it in this section.

Given centre word $c$, it maximises the value of

$$\sigma(u_c{}^{\top} v_o)$$

which matches loosely to the chance of words $c$ and $o$ co-occurring, when both of them appear, in fact, in the same context in the training corpus. Also, it minimises an identical expression when $o$ is an arbitrarily sampled term which is not available in $c$'s context.

The skip-gram model describes the context words as those restricted within a space of width $m$ around the centre word. The equation is represented as

$$\begin{aligned}\mathbb{P}(w_t \mid c) &\equiv \mathbb{P}(w_t \mid w_{t-m}, \ldots, w_{t-1}, w_{t+1}, \ldots, w_{t+m}) \\ &= \sum_{s \in S_{w_t}} \mathbb{P}(s \mid w_{t-m}, \ldots, w_{t-1}, w_{t+1}, \ldots, w_{t+m}).\end{aligned}$$

Then, the equation is re-arranged through Bayes' theorem to obtain a probability of context, given the meaning of the centre word rather than vice versa:

$$\begin{aligned}\mathbb{P}(w_t \mid c) &= \sum_{s \in S_{w_t}} \mathbb{P}(s \mid w_{t-m}, \ldots, w_{t-1}, w_{t+1}, \ldots, w_{t+m}) \\ &= \sum_{s \in S_{w_t}} \frac{\mathbb{P}(w_{t-m}, \ldots, w_{t-1}, w_{t+1}, \ldots, w_{t+m} \mid s)\, \mathbb{P}(s)}{\mathbb{P}(w_{t-m}, \ldots, w_{t-1}, w_{t+1}, \ldots, w_{t+m})}.\end{aligned}$$

Independently treating each context word:

$$\begin{aligned}\mathbb{P}(w_t \mid c) &= \sum_{s \in S_{w_t}} \mathbb{P}(s) \prod_{i=t-m, \neq t}^{t+m} \frac{\mathbb{P}(w_i \mid s)}{\mathbb{P}(w_i)} \\ &= \sum_{s \in S_{w_t}} \mathbb{P}(s) \prod_{i=t-m, \neq t}^{t+m} f(w_i, s),\end{aligned}$$

where $f(w_i,s)$ indicates the ratio between the probability of $w_i$ appearing in the context of word sense $s$, and of $w_i$ appearing in the corpus as a whole (meaning, it's unigram probability). In short, $f(w_i,s)$ provides an estimation of how many times is likely that $w_i$ appears when $s$ is present.

The convenience of the high-quality pre-trained embeddings caused by the skip-gram model (Mikolov et al., 2013c), followed by GloVe (Pennington et al., 2014), allowed WE to be successfully employed in many areas, such as parsing (Chen and Manning, 2014), sentiment analysis (Wang et al., 2016) and question answering (Seo et al., 2016), amongst others.

## 2.3 Gender Bias

Research in Social Sciences trusts linguistic analysis, since academics in Humanities know that the access to the world, culture included, is mediated by language.

Back in the 20th century, Wittgenstein et al. (1953) stated that the meaning of words is given by their applicability in everyday language. In that sense, conceptual analysis is an essential part of the work of philosophers, by explaining the rules that guide the usage of certain terms.

Thus, following the definition by Crawford (2017), bias is considered a "harmful behaviour or action which can be categorised in two groups: harms of allocation and harms of representation". The first one refers to a system that allocates or withholds resources from certain groups, such as who can get a loan, an insurance and so on. The second is when a system reinforces the subordination of some groups due to their identity, e.g. race, class, gender, etc.

By thinking about those resources, Herbelot et al. (2012) defined a pending aim for Feminist Philosophy: providing more evidence of the cultural factors that facilitate gendered behaviour and associations in humans. For that matter, they proposed the production of statistical databases on gendered distribution, so they would a) provide an ongoing philosophical research with an appropriate amount of data and b) help overcome the issues linked to the selective nature of the sources a person might read and choose as relevant.

When it comes to phenomena that relates to human self-understanding, discourse implies more than a mirror of senses. Rather, it contributes in the making of real identities. It provides the assets for people to mould their self-knowledge from and model their behaviour afterwards. This is exactly why philosophers talk about the 'power' of language in its everyday usage (Herbelot et al., 2012).

If distributions are a valuable approximation of lexical meaning, then it would be expected that the phrasal distribution of "black woman" differs from its compositional counterpart[1], for instance. Such a relevant distinction would also have linguistic relevance, as it would imply the need to take phrasal distribution into account whilst "computing meaning". In such way, both models contribute to a complete and accurate representation.

[1] According to Herbelot et al. (2012), all distributional compositionality models assume that the composition of two terms returns a distribution (lexical meaning), constituted by the sum of the two individual meanings. Both differ on how those models should be perceived. The **phrasal models** emulate the distribution of the resulting phrase itself in a large corpus. Meanwhile, the **intersective models** try to produce an adequate representation of the semantic intersection of the phrase components.
They considered the example of *big city*. Seen through a phrasal approach, it would assume that the reader associates the text with concepts like *loud*, *tube*, *light* and *crowd*, amongst others. Using an intersective model, on the other hand, it only means that a big city is a city which is big.

To recapitulate the points explained in previous pages, computers process text and the way to do so is through algorithms. A simple way to put is that every word is represented by a list of numbers called *word embeddings* (WE), which includes values regarding how words mean, how they are used, etc. Then, these figures are interpreted for every concept through a training process applied to millions of data sets of text, where words "live" in their "normal" environment.

An alternative way to see it is as if WE were coordinates in a plane, given that they are actual numbers. The distance between terms –specifically, the angle between them– measures their semantic similarity. Then, those relations can be translated to create different analogies.

The produced distributions from Computational Linguistics suggest a rational way to characterise relations between concepts, employing a huge amount of data that would be extremely difficult to handle otherwise.

Vectors in WE models capture semantic relations between words: the closer the vectors, the closer their corresponding meanings. WE are known for seizing relations not found through simple co-occurrence. They proved that common stereotypes are present, even if subtly, in the large corpora of training texts.

So, by examining embeddings and word lists, it is possible to estimate the strength of connection between neutral words and a social group. According to Garg et al. (2018), "a natural metric for the embedding bias is the average distance for women minus the average distance for men. If this value is negative, then the embedding more closely associates the occupation with men". This is how the authors claim that the adjective "honourable" is near from "man", whilst "submissive" is closer to "woman".

At the same time, the researchers related the dynamics of the embedding with the quantifiable changes in US society –e.g., demographic and occupation shifts. They concluded that the relation between embedding bias score and "reality", as calculated by occupation participation, is consistent over time. Besides, the occupations that possess a nearly 50-50 split in gender participation have a small embedding bias (Garg et al., 2018).

Also, they proposed to make comparative statements to study how the description of women through adjectives evolve over time. Garg et al. (2018) argue that an application for this work could determine how various narratives and portrayals of women developed and competed as the years passed. This approach seems to be much productive than the analogy analysis which is often used to expose how strongly human biases are encoded in language models.

### 2.3.1 What Is An Analogy?

In Linguistics, it is an equation of the form $A : B :: C : D$ ($A$ is to $B$ as $C$ is to $D$). Provided with the terms $A$, $B$ and $C$, the model must respond with the word that better represents $D$ in the comparison.

This kind of relations can be categorised in two levels: proportional and non-proportional. The first one is more concerned with organisation, form and morphology, so it accounts for systematic language regularities. In the second, meaning plays a larger role: it works in the lexico-semantic level, innovations in individual items and constructions (Fisher, 2019). Following this idea, the analogy tests set created by Mikolov et al. (2013a) consisted on morpho-syntactic types (*train is to trains as chair is to chairs*) and semantic types (*Canberra is to Australia as Stockholm is to Sweden*).

Depending on the case, the model is *forced* to output a *different* concept than any of the original ones. In other words, some models are not allowed to yield any word $D$ such that $D == B$, $D == A$ or $D == C$, since the code openly prevents it (Nissim et al., 2020). If this limitation is useful when all concepts of the analogy are expected to diverge, it turns into a problem when they *could* or *should* be the same.

This is why Nissim et al. (2020) argued that analogies as the most frequent tactic to solve this task are not the most apt. According to their findings, they might have yielded a distorted picture of bias in WE. In short: human biases are indeed present in WE and need to be resolved; however, analogies are not the most accurate tool to tackle them.

Nissim et al. (2020) offered a comprehensive list of authors to support their claim. For them, what is observed through the analogy task may be caused by irrelevant neighbourhood structure rather than to the vector offset that supposedly captured the actual analogy. Plus, Mikolov et al.'s 3COSADD method seemed to be unable to catch all linguistic regularities in the embeddings. Specifically, it seems like contextualised embeddings, such as Peters et al. (2018) and Devlin et al. (2018), do not considerate analogies as proper methods to evaluate their soundness for bias. This is because all terms of the above-mentioned equation are distinct in the original proportional analysis implementation (Mikolov et al., 2013a).

Having said that, there are other alternatives to detect the biases, such as the revised 3COSMUL method (Levy and Goldberg, 2014) and the strategy created by Bolukbasi et al. (2016). Both aim at a different take on the analogy construction, without altering the results by posing subjective inputs (Nissim et al., 2020).

Bolukbasi et. al (2016) tried to understand the biases in WE. For instance, words that are closer to *she* than to *he*. Moreover, they wanted to see the extent to which these vectors agreed with human assumptions of gender stereotypes. For this purpose, they evaluated occupation stereotypes and analogies obtained by embeddings in comparison with people, as shown in **Figure 1**.

| | def. | stereo. | | def. | stereo. |
|---|---|---|---|---|---|
| $\overrightarrow{\text{she}}-\overrightarrow{\text{he}}$ | 92% | 89% | $\overrightarrow{\text{daughter}}-\overrightarrow{\text{son}}$ | 93% | 91% |
| $\overrightarrow{\text{her}}-\overrightarrow{\text{his}}$ | 84% | 87% | $\overrightarrow{\text{mother}}-\overrightarrow{\text{father}}$ | 91% | 85% |
| $\overrightarrow{\text{woman}}-\overrightarrow{\text{man}}$ | 90% | 83% | $\overrightarrow{\text{gal}}-\overrightarrow{\text{guy}}$ | 85% | 85% |
| $\overrightarrow{\text{Mary}}-\overrightarrow{\text{John}}$ | 75% | 87% | $\overrightarrow{\text{girl}}-\overrightarrow{\text{boy}}$ | 90% | 86% |
| $\overrightarrow{\text{herself}}-\overrightarrow{\text{himself}}$ | 93% | 89% | $\overrightarrow{\text{female}}-\overrightarrow{\text{male}}$ | 84% | 75% |

***Figure 1***.*Ten possible word pairs to define gender, ordered by word frequency, along with agreement with two sets of 100 words solicited from the crowd, one with definitional and one with stereotypical gender associations.* [2]

According to these scientists, bias in WE merely reflects bias in society. However, due to the potential dangers of having machine learning systems that amplify gender stereotypes and discrimination, they recommend the wide use of neutrality as much as possible.

The Implicit Association Tests (IAT) are used in Psychology to measure subconscious gender bias in humans. Caliskan et al. (2017) adopted their core ideas, in order to measure gender bias through the difference in strength of association of concepts. In the end, the authors developed their own systems, the Word Embedding Association Test (WEAT) and Word Embedding Factual Association Test (WEFAT). They found that human bias in that kind of adapted psychological test also exists in GloVe and word2vec embeddings. Also, they demonstrated a positive correlation between the strength of association of an occupation word embedding with the female gender and the percentage of women in the same professional field.

Per Caliskan et al.'s findings, "bias should be the expected result whenever even an unbiased algorithm is used to derive regularities from any data". In other words, bias is a pattern in knowledge or information; which means that it cannot be totally removed from an NLP system.

[2] For each set of words, comprised of the most frequent 50 female and 50 male crowd suggestions, the accuracy is shown for the corresponding gender classifier based on which word is closer to a target word, e.g., the *she*-*he* classifier predicts a word is female if it is closer to *she* than *he*. With roughly 80-90% accuracy, the gender pairs predict the gender of both stereotypes and definitionally gendered words solicited from the crowd.
**Source:** Bolukbasi et al. (2016)

In the case of WEFAT, it proved that target concepts like occupation and its properties from the factual world are related to said concepts. The factual data was obtained from the U.S. Bureau of Labour Statistics, for the same year (2015) as the text corpora they were working on. The released data showed occupational categories and the proportion of women that had certain occupations under these categories. Implicit gender-occupation biases were linked to gender gaps in occupational participation.

That is to say, WEFAT contributed to the analysis of word embeddings and proved that there is empirical information about human reality encoded in language. Therefore, all implicit human biases were perpetuated in the statistical properties of the language. Moreover, historical injustice can be mirrored in training data through bias. In this sense, bias is meaning. Therefore, if an intelligent system integrated and produced this information, it would also acquire historic cultural associations, some of which can be prejudiced outcomes.

After analysing the findings from Bolukbasi et al. (2016), Zhao et al. (2018) proposed the Gender-Neutral Global Vectors (GN-GloVe) for training embedding models with protected attributes (e.g. gender).

According to their point of view, it was the first method to learn WE with protected attributes. It works by capturing them in certain dimensions and ameliorating their interpretability in representation. Also, they demonstrated that their process effectively isolated them whilst preserving the word proximity at the same time. In the end, compared with GloVe, GN-GloVe reduced the bias by 35% (Zhao et al. 2018).

The word similarity tasks assessed how well a word embedding model understood the similarity between two words compared to human-annotated rating scores. In this ground, GN-Glove achieved a better accuracy than the early studies completed before them.

The investigators concluded that their method could be applied to any language, as long a list of gender-defined words could be provided as seed words (gender pronouns, for instance). As a recommendation, they suggested to extend their approach to model other properties such as sentiments, as well as generalising their analysis beyond binary gender.

As previously shown, the work done to reduce the gender bias in word embeddings up to 2018 was either as a post-processing step (Bolukbasi et al., 2016) or as a training procedure (Zhao et al., 2018). Meanwhile, Gonen & Goldberg (2019), demonstrated that even when dramatically reducing the gender bias, as done in earlier research, it was still reflected in the geometry of the representation of “gender neutral” words, and many of the biased information could be retrieved.

Their key finding was that, nearly by representation, most word pairs kept their past similarity, even if there was a change in the axis of the gender direction. As a consequence, most words that had a specific bias before were still grouped together. Apart from adjustments in regards to certain "gendered" words, the WE's spatial geometry remained quite the same. This was a direct critic towards Bolukbasi et al. (2016), since both authors thought that they did not quantify the extensiveness before and after debiasing. Furthermore, they treated it almost as a nuance and that they did not provide any methods to deal with it either.

The resulted clustering of gendered words revealed that whilst it there is no straightforward manner to "observe" the bias, it is still manifested by the words that are *socially-marked* as feminine or masculine. For example, "nurse" will no longer be designated as an explicitly "feminine" concept, but will still be close to other "women-like" terms, such as "receptionist", "caregiver" or "teacher".

Under those circumstances, Gonen & Goldberg (2019) suggested a new mechanism for tackling this issue: the percentage of male/female socially-biased words amongst the k-nearest neighbour of the target word. While the social bias associated with a concept could not be seen straight in the new embeddings, it could be approximated by employing the gender-direction in non-debiased embeddings. That is to say, the implicit gender of words with prevalent previous bias was easy to predict based on their vectors alone.

Algorithmic discrimination is more likely to occur by linking one implicitly gendered term with other implicitly gendered words, or selecting gender-specific patterns in the corpus, by learning to condition on gender-biased words. An example could be a résumé classifier that favours male over female candidates based on commonplace cues in an existing data set, despite being "unacquainted" to gender.

### 2.3.2 Gender Bias in Language Models

During the course of this research project, it will be assumed that every statistical representation should reflect the inequality that exists in society. It is logical that such system shows this behaviour, since it is fed by that same society, namely its examples and datasets. As some authors state, it is inevitable to retain some of that bias (Prates et al., 2019)

Having said that, it is also a necessity to identify the biased language patterns, in order to recognise all forms of gender inequality. Whilst there are languages where gender neutrality is a feasible goal –e.g., English–, there are some other cases that have a longer path ahead.

On the other hand, gender-marked languages have rich grammatical structures to express ideas, both in reality and fictional environments. In order to create almost any valid sentence in this kind of language, gender is required not only in pronouns (she or he, for instance), but also in nouns, adjectives and verbs. Therefore, analysing English, with little to no gender-marking, *versus* Spanish entails a method of producing gender marking that may not have explicit evidence (Gonen & Webster, 2020).

In other words, the statistical representation of grammatical gender involves correctly assigning the grammatical gender of all entities in a sentence (Saunders et al., 2020). In some instances, this feature depends on the social gender of human referents. E.g., in the Spanish translation of the sentence "This is the doctor", "the doctor" would be either "el médico" (masculine) or "la médica" (feminine). Given that the noun refers to a certain subject, the grammatical gender enunciation should be correct for its referent.

As a starting point, it is expected in this research that, even though the distribution of gender terms may diverge from a perfect balance of 50:50, it should not diverge to the extent of misrepresenting the job occupations. As Prates et al. (2019) assumed, Google Translate demonstrates a negative gender bias, since it overestimates the frequency of male default results, even if the distribution of female employees is higher in a given professional position.

The most interesting issues here are to determine which language model exposes its preference for a particular orientation.

# 3. Methodology

As stated in previous sections, word embeddings (WE henceforth) replicate gender biases found in everyday life, since they are born from corpora that contain such prejudice.

This phenomenon is pervasive and consistent, provoking serious implications in Lexicography and Linguistics, but also in terms of Neuroscience and Law, namely Cognitive Liberty (Sommaggio, Mazzocca, Gerola & Ferro, 2004). The aim of this project is to analyse gender bias in word embeddings and language models.

The approaches outlined here have potential value for research, so WE representations are more accurate, in terms of semantic information. This is a very important step, since it is desirable to have an assessment method that is data driven, rather than based on intuitions.

In this sense, the project is proposed to discover a correlation between statistical and WE bias, as well as in adjectives in their occupational spaces. This will be done by determining the underlying reasons of their patterns, both in English and Spanish. Also, it is beneficial to find out how many professions in statistical data meet at pre-established bias. A deeper understanding of the gender preconception will be possible through the disparities amongst the various algorithms and corpora.

## 3.1 Evaluation Metrics

According to Rahutomo, Kitasuka, & Aritsugi (2012), cosine similarity is a widely implemented metric in information retrieval, which ranges between 0 to 1. The similarity between two words can be derived from calculating the cosine value between two term vectors: they are said to be similar when this value is close to 1, and different when it is close to 0. Orkphol & Yang (2019) calculated it by using their dot product and dividing it by the product of their norm, as shown in the equation below:

$$cosim(v_1, v_2) = \frac{v_1 \cdot v_2}{||v_1||||v_2||}$$

The similarity value between the query and the words is sorted from highest to lowest: the higher the similarity score, the stronger the link between them.

The best method to characterise occupations nouns in relation to their vector spaces and also in terms of adjectives is the similarity score. It adds semantic insights to the system outputs, evaluating what changes when modifying the gendered and non-gendered terms. This comes from previous research, particularly the work of Gonen & Goldberg (2019), who suggested a similar approach to the issue. Per their findings, the k-nearest neighbour of a word in a male-female semantic orientation space provided an extra set of perceptions that were more difficult to identify with the sole use of a vector plot.

For them, this is a “a natural metric for the embedding [gender] bias”. When subtracting the average distance for women minus the average distance for men, and the value turns out to be negative, then the embedding makes a closer association with masculine features. They argued that an application for this work could determine how various narratives and portrayals of women developed. Their approach seems to be much productive than the analogy analysis which is often used to expose how strongly human biases are encoded in language models.

As previously stated, two main lexicographic resources where employed in order to avoid any lack of accuracy in the classification of nouns and adjectives. Those two are the Essential British Dictionary from Cambridge University Press and the Spanish Language Dictionary from the Royal Spanish Academy (RAE).

Even though the original impression was to automatically extract the adjectives using the Natural Language Toolkit, often known as NLTK[3], the practice showed that it was not precise enough. Whilst revising the results provided by that system versus the cross-checking with dictionaries proved that the approach had to be reformulated.

Next, **Table 21** will provide the information regarding the English experiments, whilst **Table 22** shows the obtained results for Spanish.

However, when comparing language models applied on different test sets, this system alone is not valid. This is because all word spaces vary according to the model and corpus that are put into operation. Dividing the results into quartiles has a better interpretability, since all results are compatible amongst each other.

[3] NLTK is a free, open source, community-driven platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenisation, stemming, tagging, parsing, and semantic reasoning.

Its documentation describes it as "suitable for linguists, engineers, students, educators, researchers, and industry users alike". NLTK is available for Windows, Mac OS X, and Linux.

Further information is available at: https://www.nltk.org/

### 3.2 Language Models

For the sake of consistency, all the exercises used 300-dimension word embeddings in English and Spanish.

For English, the main source was the NLPL word embeddings repository[4], by the Language Technology Group at the University of Oslo. From there, for fastText and GloVe, these corpora were used: English Wikipedia Dump of February 2017, Gigaword 5th Edition and English Wikipedia Dump of February 2017 + Gigaword 5th Edition. Additionally, the Google News dataset was utilised for word2vec and GloVe.

For Spanish, the Spanish Billion Words Corpus[5] was implemented for each of the three models.

| Language | Corpus/Corpora | Number of vectors | Generation model | Number of dimensions |
|---|---|---|---|---|
| English | English Wikipedia Dump of February 2017 | 273,930 | fastText Skipgram | 300 |
| English | Gigaword 5th Edition | 262,269 | fastText Skipgram | 300 |
| English | English Wikipedia Dump of February 2017 + Gigaword 5th Edition | 260,073 | fastText Skipgram | 300 |
| English | English Wikipedia Dump of February 2017 | 273,930 | GloVe | 300 |
| English | Gigaword 5th Edition | 262,269 | GloVe | 300 |
| English | English Wikipedia Dump of February 2017 + Gigaword 5th Edition | 260,073 | GloVe | 300 |
| English | Google News Dataset | 3,000,000+ | word2vec | 300 |
| Spanish | Spanish Billion Word Corpus | 855,380 | fastText | 300 |
| Spanish | Spanish Billion Word Corpus | 855,380 | GloVe | 300 |
| Spanish | Spanish Billion Word Corpus | 1,000,653 | word2vec | 300 |

***Table 1.*** *List of analysed corpora and their number of vectors, generation models and number of dimensions.*

[4] Available at http://vectors.nlpl.eu/repository. Last accessed August 2021.
[5] Available at https://github.com/dccuchile/spanish-word-embeddings. Last accessed August 2021.

## 3.3 Analysed Occupations

Through the information of the Labor Force Statistics from the Current Population Survey (CPS) of the U.S. Bureau of Labor Statistics[6], it was possible to create two lists of the most extreme professions in the job market, along with a neutral occupation list. In order to calculate them, it was seen how many men and women were employed in each case, according to the above-mentioned survey. Then, the percentages of men/women were estimated. The same procedure was followed for the Spanish case, through the Labour Force Survey (LFS) of the National Statistics Institute[7].

Later on, the possibility of naming the occupation with just one word appeared as an issue. For instance, even though the category of "preschool and kindergarten teachers" in English had a 98.8% of women as workforce, there was not a single term that could englobe the whole meaning, so it was avoided. On the other hand, the neutral occupations were selected using those that had an estimated percentage distribution close to 50% men and 50% women.

The results of the English version are available in **Table 2**, whilst the Spanish version is in **Table 3**, located in the Findings section. As for the automatically-extracted occupations, the methodology is described in the following subsections.

## 3.4 Bias in Semantic Orientation

According to Hatzivassiloglou & McKeown (1997), the *semantic orientation*, also known as *polarity* of a word, indicates the direction in which the term deviates from the norm for its semantic group or *lexical field*. For instance, if two words relate to the same property (e. g., members of the same scalar group such as *hot* and *cold*), but demonstrate different polarities, it can be inferred that they are antonyms. In that sense, identifying the semantic orientation of words would allow a system to further improve the retrieved semantic similarity relationships in terms of gender bias.

---

[6] Available at https://www.bls.gov/cps/cpsaat11.htm. Last accessed: June 2021.
The CPS, also known as to as the household survey, is a monthly sample survey of 60,000 eligible households conducted by the U.S. Census Bureau for the Bureau of Labor Statistics, using a combination of live telephone and in-person interviews with household respondents.
It determines the demographic characteristics of people in the household and information to verify whether they are employed, unemployed, or not in the labour force.

[7] Available at https://www.ine.es/jaxiT3/Tabla.htm?t=4128. Last accessed: June 2021.
According to their website, the LFS is a continuous, quarterly survey aimed at families, the main purpose of which is to obtain data on the labour force and its various categories (employed, unemployed), as well as the population outside the labour market (inactive). The initial sample is around 65,000 households per quarter, equivalent to approximately 160,000 people.

To do so, the semantic orientation method of Turney & Littman (2003) was used, in order to score words along a semantic dimension. It works from a pair of small seed sets of words that represent two opposing points on that dimension. In this case, there were two couples of seeds. It was made by separating gender-specific terms into masculine (guy, boy, male, man) and their feminine counterparts (gal, girl, female, woman).

In the case of the **Analysed Occupations**, given that their semantic orientation results are not comparable in different vector spaces, a way to overcome it is by seeing in which quartile every occupation is located per gender, corpus, language and language model. Four types of bias qualities were defined for all of them, which were determined by applying the quartile measure to ranked semantic orientation results. Each quartile contained the frequencies that summed to the 25% of the total frequency of words. Those were:

- *Extremely High Bias*: The one that contained dominant features of a specific gender with higher rank in the first quartile. The sum of these features was maximum.
- *High Bias*: It is the one which entailed next dominant features whose rank was less than the first one, so they are classified in this particular quartile. The sum of these features was less than the first one.
- *Medium Bias*: It was the one that had next features, whose rank was less than the second one, so they all laid in the third quartile.
- *Low Bias*: It is the one that contained all the features whose ranks were less than the previous one. The sum of these weights was least and insignificant.

Using this technique, one can perceive how the diverse words are distributed and whether or not they share the same rate of gender bias. This makes it easy to make an appraisal, not only with words in one language, but in several others as well.

An alternate way of analysing semantic orientation is by creating a list of automatically extracted professions using the two sets of seeds mentioned in previous paragraphs. This inventory of terms could also work as a way to evaluate the disparity between real jobs and computer-extracted ones or how similar the bias behaves in both society and virtuality. Thus, it can be a base for future research in the field. The tables for both English and Spanish are available in **Table 5** and **Table 6**, as well as at the end of this project, as a part of the appendices.

### 3.5 Bias in Neighbour Adjectives

Once the occupation nouns were defined, as described in Analysed Occupations and Semantic Orientation of Words, it was time to find a way to add further semantic value to the results. The answer was to determine which adjectives were closer to those occupations. Why the selection, though? As Trask (1996) states:

> Canonical adjectives typically have meanings expressing permanent or temporary attributes, such as *big*, *old*, *green*, *happy* and *dry* […] Among the grammatical characteristics often displayed by adjectives are attributive position (*a big house*), predicate position (*That house is big*), comparison (*bigger, biggest*), and inflection for gender, number and case.

For every individual occupation noun, the possible answers were obtained from the 5 highest scores in the first 1,000 words of the automatically-extracted lists. This follows the research carried out by Gonen & Goldberg (2019), where they suggested that the percentage of male/female socially-biased words amongst the k-nearest neighbour of the target word should be seen a new mechanism for tackling gender bias. That way, the implicit gender of words with prevalent previous bias was easier to predict, rather than using their vectors alone.

To avoid any miscalculations whilst defining the adjectives, two dictionaries were utilised: the Essential British Dictionary from Cambridge University Press and the Spanish Language Dictionary from the Royal Spanish Academy (RAE).

Originally, it was thought that this extraction could be automatically done through NLTK. However, the revision of the results compared to the grammatical descriptions in those lexicographical resources demonstrated that it was not the right approach, so it had to be reassessed manually.

It is important to highlight that masculine, feminine and gender-neutral terms were applied, when available, since the author believed this was the only way to truly weigh the results without biasing them at the same time. That way, there could be a real comparison amongst the system outputs, evaluating what changes when modifying gendered terms. To better exemplify this point, there is the case of “businesswoman”/“businessman”/“businessperson” in English and “presentadora”/“presentador” in Spanish.

# 4. Findings

In the following pages, the author will identify and validate the semantic orientation of words that are usually related to male and female occupations, both in real life and in the static models. This will also include values for neutral words, as a way to recognise the differences with extreme professions. Then, there will be a characterisation of the adjectives that go along the way with the occupation nouns, as to add more semantic insights to the findings. The aim is to obtain a deeper understanding of the way English and Spanish operate in this regard through their examination in fastText, GloVe and word2vec.

## 4.1 Statistics vs. Word Embedding

As mentioned in the previous section, specifically in the **Analysed Occupations** of the **Methodology**, in the Labor Force Statistics from the Current Population Survey of the U.S. Bureau of Labor Statistics there is a comprehensive list of professions and occupations segmented by different variables, including gender. Using this information as a basis, it was possible to calculate the distribution of the different categories, according to the stat of those in their respective professions.

The results of the English version are available in **Table 2**.

| **Datasets** | | |
|---|---|---|
| Extreme *woman* occupations | Extreme *man* occupations | Neutral occupations[8] |
| nutritionist (91.4%) | mason (98.5%) | judge (53.9%) |
| hairdresser (90.8%) | plumber (97.7%) | artist (53.5%) |
| secretary (90.7%) | mechanic (96.7%) | photographer (52.1%) |
| nurse (88.7%) | roofer (96.7%) | packager (51.4%) |
| receptionist (88.3%) | millwright (95.8%) | dispatcher (50.8%) |
| bookkeeper (87.3%) | firefighter (95.6%) | statistician (50.3%) |
| caretaker (78.8%) | carpenter (94.5%) | bartender (49.3%) |
| librarian (77.6%) | pilot (94.4%) | scientist (45.9%) |
| veterinarian (76.3%) | constructor (94.1%) | producer (45.2%) |
| aide (75.3%) | repairer (92.5%) | coach (44.8%) |

***Table 2.*** *List of extreme and neutral occupations in English, self-made with the information of the U.S. Bureau of Labor Statistics.*

[8] The percentage rates represent the number of women as proportion of the workforce in that particular job.

In the case of the National Statistics Institute (INE) of Spain, the results were as follows:

| Datasets | | |
|---|---|---|
| Extreme *woman* occupations[9] | Extreme *man* occupations | Neutral occupations[10] |
| asistente (88.9%) | constructor (90.1%) | financiera (52.8%) |
| costurera (75%) | carpintero (80.4%) | comerciante (52.9%) |
| bibliotecaria (75%) | ingeniero (77.1%) | científico (54.9%) |
| archivista (75%) | pescador (75%) | farmaceuta (50%) |
| educadora (71.2%) | zapatero (75%) | editor (50%) |
| veterinaria (66.7%) | ebanista (75%) | presentador (50%) |
| jardinera (65.6%) | reparador (74%) | publicista (50%) |
| auxiliar (64.3%) | agricultor (73.4%) | funcionario (48.2%) |
| gestora (57.1%) | ganadero (73.4%) | artista (38.7%) |
| investigadora (57.1%) | informático (59.3%) | |

***Table 3.*** *List of extreme and neutral occupations in Spanish, self-made with the information of the National Statistics Institute.*

As a meticulous reader might see, there is one relevant thing that cannot go unnoticed. A closer look to both datasets reveals how bias is manifested in the professional field. In particular, the case of “librarian”-“bibliotecaria” is especially interesting. Apart from being included in both female-related lists, they share a very similar distribution (77.6% [US] vs. 75% [Spain]). The same could be said in “constructor”-“constructor”, where the distribution is similar but less than the female counterpart (94.1% vs. 90.1%). For the gender-neutral occupations, the distances are further apart. However, “artist”-“artista” and “scientist”-“científico” appear in both datasets.

Another aspect that should be noted is the way in which masculine-related jobs in English are extremely biased themselves. For instance, the lowest percentage for that list is “repairer” (92.5%), whilst the top feminine-related group starts with a lower score 91.4% for “nutritionist”.

#### 4.1.1. Semantic orientation in extreme and gender-neutral occupations in English

There is a second set of occupations, made out of automatically-extracted jobs from the language models through semantic orientation. The aim is to compare how these sets differ between one another, as to further explain how statistical occupations differ from the ones extracted by a word embedding.

**Table 4** and **Table 5** show the manually-obtained results. An important thing to note is that the N/A values of these tables mean that there were not enough professions in the first 1,000 words provided by the language model, so their place remained empty.

[9] Since Spanish is a gender-marked language, *woman* occupations were written in feminine.
[10] The percentage rates represent the number of women as proportion of the workforce in that particular job.
**N. B.:** There were not enough jobs or occupations in Spain that were equally distributed, so only those 9 joined the experiments.

| Extreme Feminine Professions in Word Embeddings | | | | | | | |
|---|---|---|---|---|---|---|---|
| fastText | | | GloVe | | | | word2vec |
| **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| chanteuse (-0.6311) | abbess (-0.4747) | model (-0.6722) | actress (-0.6563) | deaconess (-0.4056) | chairperson (-0.5282) | nurse-midwife (-0.4816) | chanteuse (-0.7899) |
| abbess (-0.6141) | model (-0.4502) | chairperson (-0.6672) | vedette (-0.6274) | alumna (-0.4006) | actress (-0.4793) | alumna (-0.4713) | songstress (-0.7164) |
| alumna (-0.6107) | actress (-0.4383) | supermodel (-0.6498) | chairperson (-0.6166) | mayoress (-0.3833) | dominatrix (-0.4481) | actress (-0.4365) | housewife (-0.7153) |
| vedette (-0.6011) | songstress (-0.4216) | poetess (-0.6159) | poetess (-0.5731) | abbess (-0.3781) | nurse-midwife (-0.4390) | model (-0.4360) | actress (-0.7151) |
| comedienne (-0.5925) | alumna (-0.4207) | actress (-0.5663) | alumna (-0.5647) | diva (-0.3377) | supermodel (-0.4368) | singer (-0.4334) | alumna (-0.7016) |
| sculptress (-0.5889) | huntress (-0.4200) | ex-model (-0.5228) | abbess (-0.5319) | benefactress (-0.3293) | diva (-0.4349) | benefactress (-0.4295) | comedienne (-0.7015) |
| ballerina (-0.5850) | singer (-0.4003) | general-manager (-0.5170) | ballerina (-0.5279) | nun (-0.3242) | chanteuse (-0.4335) | midwife (-0.4290) | showgirl (-0.6665) |
| songstress (-0.5808) | headmistress (-0.3956) | N/A | chanteuse (-0.5233) | nurse (-0.3204) | model (-0.4275) | hostess (-0.4066) | hostess (-0.6619) |
| mezzo-soprano (-0.5798) | N/A | N/A | patroness (-0.5141) | midwife (-0.3087) | shepherdess (-0.4218) | chanteuse (-0.3860) | Homemaker (-0.6612) |
| contralto (-0.5509) | N/A | N/A | prioress (-0.5070) | handmaid (-0.3055) | singer (-0.4145) | songstress (-0.3842) | nurse_midwife (-0.6529) |

***Table 4***. *Semantic orientation results for feminine professions in English word embeddings, manually-obtained from the first 1,000 words of every language model and corpus.*

| Extreme Masculine Professions in Word Embeddings | | | | | | | |
|---|---|---|---|---|---|---|---|
| fastText | | | GloVe | | | | word2vec |
| **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| smith (0.5915) | captain (0.5506) | quarterback (0.5766) | quarterback (0.5887) | Captain (0.4709) | groundskeeper (0.5346) | Captain (0.4998) | wizard (0.5763) |
| captain (0.5598) | cop (0.5209) | pitcher (0.5643) | player (0.5674) | shoeshine (0.4633) | wizard (0.5271) | ballplayer (0.4880) | players (0.5495) |
| manager (0.5197) | sergeant (0.5098) | actor (0.5426) | guard (0.5663) | Musketeers (0.4553) | team-owner (0.5164) | sergeant (0.4783) | N/A |
| skipper (0.5140) | thief (0.4594) | player (0.5295) | Captain (0.5477) | helmer (0.4437) | linebacker (0.4866) | linebacker (0.4705) | N/A |
| actor (0.4773) | sheriff (0.4321) | outfielder (0.5002) | officer (0.5274) | cabbie (0.4217) | quarterback (0.4284) | Colonel (0.4623) | N/A |
| player (0.4651) | Colonel (0.4317) | captain (0.4949) | coach (0.5207) | manservant (0.4196) | cornerback (0.4090) | leftfielder (0.4529) | N/A |
| coach (0.4580) | actor (0.4293) | guard (0.4884) | manager (0.5160) | cowboy (0.4126) | Manager (0.4076) | buffoon (0.4404) | N/A |
| quarterback (0.4361) | detective (0.4209) | driver (0.4871) | Colonel (0.5137) | Mafioso (0.4090) | coach-general (0.3965) | General-Major (0.4381) | N/A |
| linebacker (0.4334) | drummer (0.4166) | skipper (0.4738) | builder (0.5073) | Sergeant (0.4067) | player (0.3943) | skipper (0.4201) | N/A |
| drummer (0.4243) | policeman (0.4090) | officer (0.4595) | actor (0.4985) | Gladiator (0.4029) | N/A | centreback (0.4126) | N/A |

***Table 5.*** *Semantic orientation results for feminine professions in English word embeddings, manually-obtained from the first 1,000 words of every language model and corpus.*

| | Feminine | Masculine |
|---|---|---|
| *Total number of professions:* | 37 | 41 |
| *Number of most mentioned occupations:* | 5 | 4 |
| *Most common profession areas and total number of mentions:* | arts (7)<br>entertainment (5)<br>health (2) | defence (5)<br>sports (4)<br>arts (2) |
| *Most common professions and total number of mentions:* | actress (6)<br>alumna (6)<br>chanteuse (5)<br>abbess (4)<br>songstress (4) | captain (6)<br>actor (4)<br>player (4)<br>quarterback (4) |

***Table 6.*** *Summary of most common features in semantic orientation for feminine and masculine professions in English.*

In order to create **Table 6**, all the professions were listed. Later on, the individual terms were counted to see how many repetitions there were. For example, "actress" and "alumna" appeared 6 times each in **Table 4** and so did "captain" in **Table 5**. The *total number of professions* is the number of different occupations there were in each table. The *number of most mentioned occupations* is made by counting the jobs that were repeated more than 4 times.

As one can note, there is a higher number of masculine-related professions than feminine-related jobs. However, there is a higher representation of most mentioned occupations in the feminine examples than in the masculine cases. Plus, the professional areas linked to women have a stronger relation with regards to their men counterparts: whilst *arts* and *entertainment* account for 12 mentions, *defence* and *sports* reach just 9 references. This could be related to the high number of N/A's from word2vec.

This even applies for the most common professions, where "actress", "chanteuse" and "songstress" belong to *arts*. On the other hand, "captain", "player" and "quarterback" can be classified as *sport*-related, whilst "actor" fits in *arts*. At the same time, it is worth noticing that both "actress" and "actor" are repeated nouns for women and men, respectively. This is an important finding, since it could mean that these language model consider them as an occupation that pertains to both genders.

In this sense, "actress" appears in the original versions of GloVe and fastText, as well as in the Gigawords and the Wikipedia + Gigawords corpora. The most interesting fact is that it tends to have higher scores in both Gigawords corpora, and losses some standing when combined with

Wikipedia. This can make sense, since probably the incorporation of more vocabulary affects its performance.

Something similar occurs with "abbess". It is available in the Wikipedia corpora (fastText and GloVe), ranks as number 1 in Gigawords for fastText and then vanishes altogether.

As per "songstress", its behaviour is quite contradictory. In fastText, it appears in the Wikipedia and the Gigawords corpora, but it does not in the combination of both. The opposite is true in GloVe: one cannot find it in the Wikipedia and the Gigawords corpora, but it is present when both corpora are combined.

This strange conduct is not exclusive to feminine-related nouns. In the case of "captain", it holds the 2nd and 3rd spot in the Wikipedia and Gigawords for fastText, respectively. However, when combining both of them, the ranking alters and its standing falls to the 5th position. For GloVe, it holds the 1st place for the Wikipedia and the Wikipedia + Gigawords corpora, even though it is not present in the Gigawords corpus.

The opposite happens in "actor". It appears as 5th and 7th in Wikipedia and Gigawords for fastText (respectively), but then goes up to 3rd position in the combination of both corpora.

On the other hand, in regards to the feminine professions, there are two evident things. First, the high number of words ending in *-ess*. Those account for a total of 14 different nouns ("abbess", "actress", "benefactress", "deaconess" … etc.), some of them being the most common professions, as mentioned in previous paragraphs. Also, the presence of foreign words: 4 in Italian ("ballerina", "contralto", "diva", "mezzo-soprano"), 3 in French ("chanteuse", "comedienne" and "vedette") and 1 in Latin ("alumna").

Linked to the *-ess* ending, it is a simple way to determine the femininity of a word. According to Harper (n.d.), it comes from the Middle English *-esse*, which in turn was derived from the Old French *-esse*, from Late Latin *-issa* and from Ancient Greek -ισσα (*-issa*). In English, words of this kind were adopted (e.g.: *countess*, *duchess*, *mistress* and *princess*) or formed nouns in *-er* (*enchantress* and *sorceress*). When this suffix is added to a noun ending in *-tor*, *-ter*, the vowel before *r* is commonly elided, like the case of *actress* (*actor* + *-ess*).

Similarly, this could also be the case for the foreign words. Since all of them come from Romance languages, they include gender in themselves, so it would simple for a machine to learn them and then make out of relation with female seeds.

In the case of masculine occupations, there is a strong link with sports, as already mentioned. Particularly, there is a heavy weight for American football ("centreback", "cornerback", “linebacker” and “quarterback”) and baseball (“outfielder”, “pitcher”). This could happen because of the corpora’s origin. They all come from American sources, therefore, it is not surprise to see this.

The nature of the corpora could also be the reason for the following finding. There are two sets of 7 nouns each that contain occupations from the past, because they are not common in the 21$^{st}$ century. It happens in feminine (“abbess”, “benefactress”, “deaconess”, “huntress”, “patroness”, “prioress”, “nun”) and masculine (“buffoon”, “gladiator”, “manservant”, “musketeer”, “shoeshine”, “smith”, “wizard”). The main difference between them is that words related to women are more spiritual, whilst the corresponding terms for men lean more towards servitude-like concepts.

From a total of 14 examples, 5 masculine words (“gladiator”, “manservant”, “musketeer”, “shoeshine” and “smith”), as well as 4 feminine terms (“abbess”, “benefactress”, “deaconess”, and “nun”) come up as the result of the Wikipedia corpus. The cited situation was apparent from afar, since the extreme female occupations from the Wikipedia corpus in fastText had already demonstrated a clear tendency towards religion.

Any curious reader would not stop thinking how is it that such features carry on in the present, even when those terms seem to belong to a long past time. According to Stalin (1977), language comes as the sum of ages and cycles that occur in a given society. This is the reason why it has an incomparably longer life than any base or any superstructure. History is witnessing the fading or even destruction of successive bases and their corresponding superstructures, and yet, this does not translate into the elimination of the current language at each moment and the birth of another with new vocabulary and grammatical structure.

For Stalin (1977), language has not been created to satisfy the needs of any one class, but of the whole society, of all social statuses. It is precisely for this, his argument concludes, that it can serve equally to an old and dying regime and to a new and rising system, by combining the old base and the new.

Even so, differences must persist. In the case at hand, there are really no coincidences between the statistical and the word embedding results, since both produce very different outcomes that do not have a reflection on the other. To further explain this reasoning, the statistical insights will be explained in the following pages.

| Extreme Feminine Professions in U.S. Labour Statistics | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | fastText | | | GloVe | | | | word2vec |
| | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| *Nutritionist* | -0.2831 | -0.3308 | -0.2529 | -0.3341 | -0.3051 | -0.3045 | -0.3378 | -0.3162 |
| *Hairdresser* | 0.1064 | -0.1763 | -0.0220 | -0.1509 | -0.0208 | -0.1646 | -0.0762 | -0.3548 |
| *Secretary* | -0.0716 | -0.0673 | -0.0379 | 0.0566 | -0.0807 | 0.0234 | 0.0321 | -0.1839 |
| *Nurse* | -0.3204 | -0.2796 | -0.2256 | -0.2450 | -0.1942 | -0.1612 | -0.1685 | -0.5054 |
| *Receptionist* | -0.0155 | -0.2426 | -0.1177 | -0.2731 | -0.1065 | -0.3315 | -0.1828 | -0.4995 |
| *Bookkeeper* | 0.0987 | -0.0670 | 0.1112 | -0.2203 | 0.0214 | -0.1229 | -0.1575 | -0.3307 |
| *Caretaker* | 0.0381 | 0.0628 | 0.0898 | 0.1944 | 0.0569 | 0.1352 | 0.1064 | 0.1816 |
| *Librarian* | -0.1959 | -0.2609 | -0.2298 | -0.2715 | -0.1394 | -0.2752 | -0.2586 | -0.4740 |
| *Veterinarian* | 0.0300 | -0.0752 | 0.0141 | -0.0911 | 0.0240 | 0.0195 | -0.0036 | -0.1031 |
| *Aide* | 0.1747 | 0.1588 | 0.2828 | 0.1152 | 0.1215 | 0.2605 | 0.3359 | -0.0090 |
| *Low Bias Limit (fem)* | 0 – –0.0287 | 0 – -0.0305 | 0 – -0.0303 | 0 – -0.0729 | 0 – -0.0502 | 0 – -0.0761 | 0 – -0.0764 | 0 – -0.0423 |
| *Medium Bias Limit (fem)* | -0.0288 – -0.0623 | -0.0306 – -0.0683 | -0.0304 – -0.0663 | -0.0730 – -0.1413 | -0.0503 – -0.1014 | -0.0762 – -0.1450 | -0.0765 – -0.1437 | -0.0424 – -0.0891 |
| *High Bias Limit (fem)* | -0.0624 – -0.1083 | -0.0684 – -0.1268 | -0.0664 – -0.1192 | -0.1414 – -0.2231 | -0.1015 – -0.1655 | -0.1451 – -0.2263 | -0.1438 – -0.2227 | -0.0892 – -0.1546 |
| *Extremely High Bias Limit (fem)* | Less than -0.1084 | Less than -0.1269 | Less than -0.1193 | Less than -0.2232 | Less than -0.1656 | Less than -0.2264 | Less than -0.2228 | Less than -0.1547 |
| *Low Bias Limit (mas)* | 0 – 0.0461 | 0 – 0.0502 | 0 – 0.0500 | 0 – 0.0344 | 0 – 0.0337 | 0 – 0.0347 | 0 – 0.0331 | 0 – 0.0395 |
| *Medium Bias Limit (mas)* | 0.0461 – 0.0954 | 0.0503 – 0.1016 | 0.0500 – 0.1021 | 0.0345 – 0.0783 | 0.0338 – 0.0739 | 0.0348 – 0.0795 | 0.0331 – 0.0756 | 0.0396 – 0.0866 |
| *High bias Limit (mas)* | 0.0955 – 0.1602 | 0.1017 – 0.1686 | 0.1022 – 0.1710 | 0.0784 – 0.1451 | 0.0740 – 0.1315 | 0.0796 – 0.1481 | 0.0757 – 0.1433 | 0.0867 – 0.1579 |
| *Extremely High Bias Limit (mas)* | More than 0.1603 | More than 0.1687 | More than 0.1711 | More than 0.1452 | More than 0.1316 | More than 0.1482 | More than 0.1434 | More than 0.1580 |

***Table 7.*** *Semantic orientation results for extreme feminine occupations in English by quartiles, using the words obtained from the U.S. Bureau of Labor Statistics.*

As stated on previous sections, the author of this report extracted the information from the Labor Force Statistics of the Current Population Survey in the U.S. Bureau of Labor Statistics and ranked it according to the jobs or occupations mostly occupied by women.

Hence, the creation of **Table 7** consisted on using the top female-related occupations and organising them according to their semantic orientation, as computed by three different language models and several corpora. This feature, which can also be referred as polarity of a word, specifies the course in which each term deviates from the norm for its lexical field. Identifying this characteristic is also a way to understand the way language models work, since it allows to see the relationships they create in regards to gender.

But calculating the semantic orientation with different language models is not comparable, given the dissimilar vector spaces they create. So, in order, to skip this issue, it was determined to compute the total number of words in terms of semantic orientation, and then categorise them in quartiles, following the explanation made in the methodology section (**3.4 Bias in Semantic Orientation**). Those were low, medium, high and extremely high bias, both for masculine and feminine associations. This way, it is possible to perceive how all terms are allocated and how they vary in comparison.

Thus, it is clear that the patterns comply with the previous research by Bolukbasi et al. (2016). Particularly, the professions that were named as “extreme she occupations” by those scholars appear here as well: “nurse”, “receptionist”, “librarian”, “hairdresser” and “bookkeeper”.

In this specific exercise, though, there are certain aspects that are worth to be mentioned. First of all, three occupations coincide both in language models and statistical distribution. The absolute champion is “nutritionist” which not only ranks as high in the “real professions” division, but also in fastText, GloVe and word2vec. There is complete unanimity in all of them, since their numbers always provide extremely high feminine scores.

Next, comes the case of “librarian”. It is regarded 7 times as extremely high biased towards feminine values, and once as highly biased in the same direction.

In the third position, there is the case of “nurse”, which not only ranks high in earlier investigations, but also demonstrates that it also has a high score in reality.

On the other hand, the word “aide” appears in the classification with extremely high or high masculine scores in 6 out of 8 results, even though reality conflicts with this notion.

For “caretaker”, there are 3 mentions in the same way (1 extremely high and 2 high). Also, all of its values correspond to masculine-related values. It reaches a high biased result in word2vec, which, by the way, has a very strong tendency to providing feminine results. In this example, it does not matter that the word itself could be identified as a job of “caring” for others, a trait that is often assigned to women.

As Bolukbasi et al. (2016) showed, word embeddings may be capable of circumventing the report of bias. This situation occurs because these models are trained using methods that require large amounts of data to extract associations and relationships. That is why the results may be altered according to the kind of data that is fed into the system. So, in order to compare the results amongst language models, different corpora were utilised. Further information can be found in **Table 1**.

In regards to the extreme feminine occupations, some of the results in the language models coincide with the socially established bias for labour. For instance, in the case of word2vec, 8 out of 10 professions were considered as highly or extremely high in terms of their quartile results. These outcomes are followed by the original corpus of GloVe and the Gigawords corpus analysed with the fastText configuration (6 each).

It is worth noticing the behaviour of the corpora: Wikipedia provides the same number of high and extreme high scores in fastText and GloVe (4 each). Also, there is a decrease of bias in the Gigawords corpus from fastText (6) to GloVe (5), but an increase in the Wikipedia + Gigawords corpus, following the same direction (3 vs. 5).

This goes hand to hand with the erratic behaviour demonstrated in computer-generated lists of extreme occupations. So, the lack of consistency shown by the language models keeps accumulating during this research.

On the other hand, it must be highlighted that exists some results that show that the occupational classification for feminine-linked professions is not monolithic. They range from the 3 examples of high or extremely high masculine-related scores in Wikipedia (fastText) to at least one sample in every other corpus.

| | fastText | | | GloVe | | | | word2vec |
|---|---|---|---|---|---|---|---|---|
| **Extreme Masculine Professions in U.S. Labour Statistics** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| *Mason* | 0.2626 | 0.2933 | 0.2787 | 0.2797 | 0.1322 | 0.2036 | 0.2455 | 0.3117 |
| *Plumber* | 0.3320 | 0.2643 | 0.2975 | 0.1124 | 0.1667 | 0.2818 | 0.2517 | 0.2395 |
| *Mechanic* | 0.2279 | 0.4261 | 0.3082 | 0.2637 | 0.2101 | 0.4331 | 0.4077 | 0.3289 |
| *Roofer* | 0.2865 | 0.2174 | 0.3156 | 0.0886 | 0.1821 | 0.1478 | 0.1063 | 0.3223 |
| *Millwright* | 0.2468 | N/A | 0.2422 | -0.0550 | 0.1417 | N/A | 0.1100 | 0.0755 |
| *Firefighter* | 0.1878 | 0.2065 | 0.2703 | 0.1543 | 0.2674 | 0.3207 | 0.3322 | 0.2186 |
| *Carpenter* | 0.2677 | 0.3104 | 0.3423 | 0.3455 | 0.3159 | 0.2725 | 0.4058 | 0.3582 |
| *Pilot* | 0.1141 | 0.0870 | 0.2096 | 0.2064 | 0.2995 | 0.2895 | 0.3818 | 0.0708 |
| *Constructor* | 0.0186 | 0.0584 | 0.0950 | -0.0588 | 0.0192 | -0.0645 | 0.0334 | -0.0330 |
| *Repairer* | 0.2835 | 0.2164 | 0.2568 | 0.1159 | 0.2399 | 0.1118 | 0.0646 | 0.0948 |
| *Low Bias Limit (fem)* | 0 – –0.0287 | 0 – -0.0305 | 0 – -0.0303 | 0 – -0.0729 | 0 – -0.0502 | 0 – -0.0761 | 0 – -0.0764 | 0 – -0.0423 |
| *Medium Bias Limit (fem)* | -0.0288 – -0.0623 | -0.0306 – -0.0683 | -0.0304 – -0.0663 | -0.0730 – -0.1413 | -0.0503 – -0.1014 | -0.0762 – -0.1450 | -0.0765 – -0.1437 | -0.0424 – -0.0891 |
| *High Bias Limit (fem)* | -0.0624 – -0.1083 | -0.0684 – -0.1268 | -0.0664 – -0.1192 | -0.1414 – -0.2231 | -0.1015 – -0.1655 | -0.1451 – -0.2263 | -0.1438 – -0.2227 | -0.0892 – -0.1546 |
| *Extremely High Bias Limit (fem)* | Less than -0.1084 | Less than -0.1269 | Less than -0.1193 | Less than -0.2232 | Less than -0.1656 | Less than -0.2264 | Less than -0.2228 | Less than -0.1547 |
| *Low Bias Limit (mas)* | 0 – 0.0461 | 0 – 0.0502 | 0 – 0.0500 | 0 – 0.0344 | 0 – 0.0337 | 0 – 0.0347 | 0 – 0.0331 | 0 – 0.0395 |
| *Medium Bias Limit (mas)* | 0.0461 – 0.0954 | 0.0503 – 0.1016 | 0.0500 – 0.1021 | 0.0345 – 0.0783 | 0.0338 – 0.0739 | 0.0348 – 0.0795 | 0.0331 – 0.0756 | 0.0396 – 0.0866 |
| *High bias Limit (mas)* | 0.0955 – 0.1602 | 0.1017 – 0.1686 | 0.1022 – 0.1710 | 0.0784 – 0.1451 | 0.0740 – 0.1315 | 0.0796 – 0.1481 | 0.0757 – 0.1433 | 0.0867 – 0.1579 |
| *Extremely High Bias Limit (mas)* | More than 0.1603 | More than 0.1687 | More than 0.1711 | More than 0.1452 | More than 0.1316 | More than 0.1482 | More than 0.1434 | More than 0.1580 |

***Table 8.*** *Semantic orientation results for extreme masculine occupations in English by quartiles, using the words obtained from the U.S. Bureau of Labor Statistic*

In nearly all cases of masculine occupations (**Table 8**), there is a clear tendency of profession nouns to be highly regarded as extremely biased. The only exception is “constructor”, which contains 2 mentions as low masculine bias, 3 as low feminine bias and 3 more as medium masculine bias. Apart from that, “millwright” contains one value of low feminine bias and medium masculine bias in the original GloVe corpus and in word2vec, respectively. Then “pilot” has two instances of medium masculine bias, in Gigawords with fastText and word2vec.

In terms of language models, fastText provides the most biased results: 9 out of 10 professions in the Wikipedia and the Wikipedia + Gigawords corpora. It is interesting that the Gigawords corpus had only 7 biased occupations, so it did not help the combined corpus to mitigate the partiality. Then, Wikipedia calculated through GloVe also provided 9 prejudiced scores.

The portrayed situation offers a clear example of how the English-speaking people utilise their language repertoire to describe those words. This is the consequence of the incorporation of lexical usages that reflect a social reality.

As Saussure & Alonso (1945) stated, amongst all the individuals linked by a language, there is the establishment of a sort of average, where they can all reproduce the same signs associated with the equivalent concepts –to a certain extent, but not exactly equal–.

It is through the functioning of the receptive and coordinative faculties in the speaking subjects that a *mark* is formed, which causes an *impression* to all. This is how the social crystallisation of language originates.

So, if a lexicographer wanted to record these conventions in a dictionary, they would carry it out in an exercise of veracity, since it is just the reflection of actual linguistic usage and social practice. This does not mean that it would be a proselytization of the words as masculine professions or the promotion of certain attitudes or behaviours, but rather a description of what happens in current daily life.

Nonetheless, these results provide the necessary data for society itself to identify the existence of undesirable linguistic bias. Its removal must be encouraged through education.

This situation is not new, since it appears as early as Aristotle's times. In his work *Politics* (2013), the philosopher affirmed that:

> The Greek word *logos* [word] means both ‘speech’ and ‘reason’; it is a man’s reasoning ability that enables him to distinguish between the just and unjust, and therefore to conduct himself morally in relation to others in a way that makes human community possible—whether in a household or a polis.

| | Gender-Neutral Professions in Labour Statistics | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | fastText | | | GloVe | | | | word2vec |
| | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| *Judge* | -0.0474 | -0.0668 | 0.0369 | 0.0872 | 0.1172 | 0.1534 | 0.2093 | -0.0719 |
| *Artist* | 0.0135 | 0.0210 | 0.0306 | 0.1643 | 0.0687 | 0.2282 | 0.1537 | -0.0447 |
| *Photographer* | 0.0388 | 0.0350 | 0.1276 | 0.0922 | 0.0917 | 0.1379 | 0.1790 | 0.0169 |
| *Packager* | 0.0604 | -0.0140 | 0.0463 | 0.0408 | 0.1202 | -0.0816 | -0.0565 | -0.2726 |
| *Dispatcher* | 0.2042 | 0.1351 | 0.2074 | -0.0228 | 0.0899 | 0.1316 | 0.1934 | -0.0497 |
| *Statistician* | 0.0135 | 0.2044 | 0.0923 | -0.0589 | -0.0648 | 0.0242 | 0.0154 | 0.1104 |
| *Bartender* | 0.3213 | 0.1365 | 0.2132 | 0.0503 | 0.3360 | 0.1573 | 0.1915 | -0.0301 |
| *Scientist* | -0.0140 | 0.0621 | 0.0473 | 0.0712 | 0.0879 | 0.2353 | 0.2157 | -0.0157 |
| *Producer* | 0.0896 | 0.0674 | 0.0621 | 0.2088 | 0.1879 | 0.2006 | 0.1922 | -0.0892 |
| *Coach* | 0.2158 | 0.3549 | 0.4003 | 0.4580 | 0.2444 | 0.4763 | 0.5207 | 0.3775 |
| *Low Bias Limit (fem)* | 0 – –0.0287 | 0 – -0.0305 | 0 – -0.0303 | 0 – -0.0729 | 0 – -0.0502 | 0 – -0.0761 | 0 – -0.0764 | 0 – -0.0423 |
| *Medium Bias Limit (fem)* | -0.0288 – -0.0623 | -0.0306 – -0.0683 | -0.0304 – -0.0663 | -0.0730 – -0.1413 | -0.0503 – -0.1014 | -0.0762 – -0.1450 | -0.0765 – -0.1437 | -0.0424 – -0.0891 |
| *High Bias Limit (fem)* | -0.0624 – -0.1083 | -0.0684 – -0.1268 | -0.0664 – -0.1192 | -0.1414 – -0.2231 | -0.1015 – -0.1655 | -0.1451 – -0.2263 | -0.1438 – -0.2227 | -0.0892 – -0.1546 |
| *Extremely High Bias Limit (fem)* | Less than -0.1084 | Less than -0.1269 | Less than -0.1193 | Less than -0.2232 | Less than -0.1656 | Less than -0.2264 | Less than -0.2228 | Less than -0.1547 |
| *Low Bias Limit (mas)* | 0 – 0.0461 | 0 – 0.0502 | 0 – 0.0500 | 0 – 0.0344 | 0 – 0.0337 | 0 – 0.0347 | 0 – 0.0331 | 0 – 0.0395 |
| *Medium Bias Limit (mas)* | 0.0461 – 0.0954 | 0.0503 – 0.1016 | 0.0500 – 0.1021 | 0.0345 – 0.0783 | 0.0338 – 0.0739 | 0.0348 – 0.0795 | 0.0331 – 0.0756 | 0.0396 – 0.0866 |
| *High bias Limit (mas)* | 0.0955 – 0.1602 | 0.1017 – 0.1686 | 0.1022 – 0.1710 | 0.0784 – 0.1451 | 0.0740 – 0.1315 | 0.0796 – 0.1481 | 0.0757 – 0.1433 | 0.0867 – 0.1579 |
| *Extremely High Bias Limit (mas)* | More than 0.1603 | More than 0.1687 | More than 0.1711 | More than 0.1452 | More than 0.1316 | More than 0.1482 | More than 0.1434 | More than 0.1580 |

***Table 9.*** *Semantic orientation results for gender neutral occupations in English by quartiles, using the words obtained from the U.S. Bureau of Labor Statistics*

As seen in **Table 9**, there is a persistent trend in neutral professions that provides a stronger link to masculine-related values and diverse occupations. Such tendency remains in the so called "gender-neutral" professions. The most consistent case is "coach", that has 8 out of 8 results classified as extremely high bias towards masculine values. It is followed by "bartender" (6 out of 8 scores) and, finally, "dispatcher" and "photographer" (5 out of 8).

Then, most of the terms has no relation to extreme feminine values. The only two exception are "producer" that (4 masculine vs. 1 feminine) and "packager", which has the most balanced recounting: one example for each side.

In regards to the language models and corpora, there is clear path of highest bias in GloVe. Both the Wikipedia and the Gigawords corpora have the same number of listed professions: 8. Surprisingly, the Wikipedia + Gigawords corpus did not produce differences in total number of biased professions towards masculine values, so it stays in 8 as well.

Just like in the extreme masculine-related jobs, there is almost agreement in not providing female-like values for gender-neutral occupations. The only exceptions appear in word2vec, where one can find an extremely high bias score and another high bias grade: "packager" and "producer", respectively (as it was mentioned beforehand).

As a summary, in **Table 10** the reader will find how the bias is distributed per language model and corpus.

| | fastText | | | GloVe | | | | word2vec |
|---|---|---|---|---|---|---|---|---|
| | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| *Most biased results in feminine professions* | 4 | 6 | 3 | 6 | 4 | 5 | 5 | 8 |
| *Most biased results in masculine professions* | 9 | 7 | 9 | 8 | 9 | 8 | 8 | 8 |
| *Most biased results in gender-neutral professions*[11] | 3 | 4 | 4 | 5 | 8 | 8 | 8 | 2 |

***Table 10.*** *Gender bias distribution per language model and corpus in English. Self-made.*

To make this table, the data corresponding to extremely high and high bias were counted for each of the words available in the various language models studied, and then added accordingly.

[11] In this case, bias goes towards masculine-related values.

The first thing to notice is the bias allocation in the language models, which is distributed exactly in 3 subgroups for each class (masculine, feminine and gender-neutral). The way in which the different types of prejudice are represented in the table is remarkable. Even the fact that the masculine and gender-neutral terms share the direction of the bias does not mean that their allotment in regards to corpus and models is the same. The only coincidence in that sense appears in the Wikipedia corpus with the GloVe configuration.

Second, the intensity of the bias must be mentioned. According to this research, the number of biased professions tends to be higher in the case of male, gender-neutral and female occupations, in that order. Hand in hand with other discoveries made so far, feminine-related and gender-neutral occupations tend to present certain similarities in intensity, while masculine-related jobs go a go ahead.

At this point, it is worth comparing the results of the statistical extreme occupations and how they would change if the semantic orientation was its norm, as seen in **Table 11**. Analogously to the procedure followed for the creation of **Table 10**, the information of the extremely high and high bias was counted for each of the occupations available in the various language models studied, and then added.

The extreme feminine occupations lose “caretaker” and “aide” to the extreme masculine set.

The gender-neutral category mislays most of its collection, which goes into the hands of the extreme masculine category. In the end, its only remaining occupation is “packager”. However, it is not a total loss: the gender-neutral group manages to snatch “constructor” from the extreme masculine professions. This situation leaves the extreme masculine sort with almost 2 out of each 3 words.

It is extraordinary to see how similar the original extreme occupation lists resemble the results from the semantic orientation. In terms of masculine-related intensity, the distribution is very similar: even “mason”, “plumber”, “mechanic” and “roofer” retain their first spots, joined by “firefighter” and “carpenter”. “Coach” appears from the gender-neutral list and sneaks into the top places, whilst “repairer” comes from behind in the original ranking. They are followed by newly-added terms from the other two original classifications, with the exception of “pilot” and “millwright”, that already were available in the previous version of the masculine-related list.

In the case of extreme feminine words, “nutritionist”, “nurse”, “librarian” and “receptionist” are the strongest contenders, both in statistics and semantic orientation. The remaining occupations demonstrate a medium or low tendency towards this kind of values.

| Occupations' list according to their statistical distribution | | |
|---|---|---|
| Extreme *woman* occupations | Extreme *man* occupations | Neutral occupations |
| nutritionist | mason | judge |
| hairdresser | plumber | artist |
| secretary | mechanic | photographer |
| nurse | roofer | packager |
| receptionist | millwright | dispatcher |
| bookkeeper | firefighter | statistician |
| caretaker | carpenter | bartender |
| librarian | pilot | scientist |
| veterinarian | constructor | producer |
| aide | repairer | coach |
| **Occupations' list according to their semantic orientation distribution** | | |
| Extreme *woman* occupations | Extreme *man* occupations | Neutral occupations |
| nutritionist | mason | packager |
| nurse | plumber | constructor |
| librarian | mechanic | |
| receptionist | roofer | |
| hairdresser | firefighter | |
| bookkeeper | carpenter | |
| secretary | coach | |
| veterinarian | repairer | |
| | aide | |
| | pilot | |
| | bartender | |
| | photographer | |
| | dispatcher | |
| | millwright | |
| | judge | |
| | producer | |
| | caretaker | |
| | artist | |
| | scientist | |
| | statistician | |

***Table 11.*** *Comparison between labour statistics and semantic orientation distributions in English. Self-made calculations.*

So, a question arises: does an occupation appear high in the semantic orientation rank because of the composition of the workforce or does the composition of the workforce determines the semantic orientation score?

It is now proven how the representational function inevitably plays a role in the verbal realisation of language. No wonder Bühler & Marías (1950) considered it to be one of the three fundamental functions of language. For them, it complements the other two: the emotive function, through which feelings are expressed; and the conative function, which serves to influence the conscience and behaviour of others. The individual and social selves are expressed, respectively, through these last two functions. The first one, on the other hand, works to relate the speaker to their reality.

Indeed, even before textual writing fixed the language in stable and easily understandable signs, language has been accompanied by the idea that it not only reproduces reality, but also creates it, through the intervention of a "divine power".

In the book of Genesis, for instance, the creation of the world is justified in these terms. Yahweh made it possible by a purely linguistic operation, when "God said, 'Let there be light', and there was light. God saw that the light was good, and he separated the light from the darkness. God called the light 'day', and the darkness he called 'night'. And there was evening, and there was morning—the first day" (New International Version, 2011, Genesis 1:3-5).

Certain arguments concerning language bias are based on these notions. To some extent, they can be traced in sacred texts such as the one mentioned above. There, the creation is described in a mythical and theological manner. This is an approach that was later readapted by the medieval scholastic philosophy, from which it re-emerged under the name of "nominalism". Nowadays, it is also argued that language creates reality, when the truth is that words are a phenomenon of what really counts, which are the things themselves.

All this to say that words, instead of creating reality, are created themselves. They consist of a phonetic construction structured in links –sounds, syllables, words– whose coupling to a reference –the designated object, fact or reality– is based on the principle of arbitrariness.

This reasoning could be further explained by the way WE analyse different words. If there was indeed a bias towards "feminine" professions, all words would be marked in red. In the experiments, however, the closest to this scenario happens in word2vec. Not even there one can find a uniform tagging trend.

#### 4.1.2. Semantic orientation in extreme and gender-neutral occupations in Spanish

Just as it happened in the previous subsection, the first step will be to extract a set of occupations from three language models, using semantic orientation with the same corpus. Since there are already some models from English, the comparison will not be just in one language but rather from a bilingual perspective.

**Table 12** and **Table 13** show the obtained results for feminine and masculine occupations, respectively.

### Extreme Feminine Professions in Word Embeddings

| fastText | GloVe | word2vec |
|---|---|---|
| empresaria (-0.7447) | poetisa (-0.7980) | empresaria (-0.7608) |
| podóloga (-0.7031) | fotógrafa (-0.7938) | politóloga (-0.7434) |
| sinóloga (-0.6992) | ventrílocua (-0.7811) | jefa (-0.7343) |
| directora (-0.6896) | empresaria (-0.7477) | directora (-0.7338) |
| ginecóloga (-0.6876) | rejoneadora (-0.7404) | escritora (-0.7307) |
| exactriz (-0.6781) | actriz (-0.7301) | abogada (-0.7195) |
| aviadora (-0.6694) | presentadora (-0.7147) | presidenta (-0.7068) |
| escritora (-0.6676) | embajadora (-0.7123) | presentadora (-0.7037) |
| repositora (-0.6618) | bailarina (-0.7072) | actriz (-0.7016) |
| abogada (-0.6571) | camarera (-0.7059) | locutora (-0.6965) |

***Table 12.*** *Semantic orientation results for feminine professions in Spanish word embeddings, self-made.*

### Extreme Masculine Professions in Word Embeddings

| fastText | GloVe | word2vec |
|---|---|---|
| acopiador (0.4578) | cuidador (0.5702) | sabio (0.4977) |
| malhechor (0.4300) | discípulo (0.5688) | capuchino (0.4937) |
| timbalero (0.4193) | filósofo (0.5515) | excavador (0.4771) |
| oteador (0.4125) | educador (0.5030) | carpintero (0.4356) |
| discípulo (0.3953) | biólogo (0.4859) | pensador (0.4205) |
| pensador (0.5920) | confitero (0.4777) | comentador (0.4164) |
| thaumaturgo (0.3918) | investigador (0.4772) | artesano (0.4126) |
| zelador (0.3876) | antropólogo (0.4749) | cuidador (0.4042) |
| baqueano (0.3865) | sociólogo (0.4716) | discípulo (0.4023) |
| techador (0.3750) | franciscano (0.4710) | cantor (0.4015) |

***Table 13.*** *Semantic orientation results for masculine professions in Spanish word embeddings, self-made.*

| | Feminine | Masculine |
|---|---|---|
| *Total number of professions:* | 23 | 26 |
| *Number of most mentioned occupations:* | 6 | 3 |
| *Most common profession areas and total number of mentions:* | business (5)<br>arts (4)<br>entertainment (2)<br>law (2) | education (5)<br>care (2) |
| *Most common professions and total number of mentions:* | empresaria (3)<br>abogada (2)<br>actriz (2)<br>directora (2)<br>escritora (2)<br>presentadora (2) | discípulo (3)<br>cuidador (2)<br>pensador (2) |

***Table 14.*** *Summary of most common features in semantic orientation for feminine and masculine professions in Spanish.*

The creation of **Table 14**, was the same as in English: both sets of professions were listed. Then, the repeated terms were summed and extracted from them, leaving just one sample. The updated terminology register was counted and that is how the total number of professions was obtained.

A similar situation to English applies to the total number of occupations. First, there are more masculine-related profession than female-linked jobs. Even so, the proportion in regards to English stays in a very similar proportion, although Spanish has only three language models (masculine: 26 vs. 41 - feminine: 23 vs. 37). Also, as in the previous section, one can tell a higher representation of most mentioned occupations in the feminine examples than in the masculine cases.

It is worth noticing the big difference in terms of semantic orientation values. Female-related positions had values that were much higher than their male counterparts:

| | Feminine values | Masculine values |
|---|---|---|
| *fastText* | From -0.7447<br>to -0.6571 | From 0.4578<br>to 0.3750 |
| *GloVe* | From -0.7980<br>to -0.7059 | From 0.5702<br>to 0.4710 |
| *word2vec* | From -0.7608<br>to -0.6965 | From 0.4977<br>to 0.4015 |

***Table 15***. *Semantic orientation range values for feminine and masculine occupations in Spanish.*

Besides, the professional areas linked to women have a stronger relation with regards to their men counterparts: whilst *business* and *arts* account for 9 mentions, *education* and *care* reach just 7 references. This relationship is much lower than English, but the trend stays.

As for the feminine professional areas, *business* included “empresaria” and “directora”. In the case or *arts*, “actriz” and “escritora” were put together. For *entertainment* and *law*, there is one occupation for each: “presentadora” and “abogada”. In the case of their masculine counterparts, “discípulo” and “pensador” were classified as *education*, whilst “cuidador” was assigned to *care*.

It is interesting to see how women are classified as subjects meant to amuse others, either in arts or entertainment. This phenomenon also occurred in in English.

Also, it must be mentioned that there is lack of tough activities linked to men. Whilst in English there were references to *defence* and *sports*, *education* shows up in Spanish. More surprisingly is the category *care*, which quite often assigned to women, but it happens the opposite in this new analysis.

The term “empresaria” appears in the first spot in fastText and word2vec, whilst it remains the 4$^{th}$ in GloVe. The analogous behaviour of the former is maintained in “directora”, where it shows up in 4$^{th}$ position. In the case of “discípulo”, it can be found in the three language models: 2$^{nd}$ in GloVe, 5$^{th}$ in fastText and 9$^{th}$ in word2vec. Meanwhile, “pensador” is available in word2vec (5$^{th}$) and fastText (6$^{th}$).

In terms of etymology, there are some aspects that should be considered. There are at least 5 suffixes that are present in the Spanish occupations:

- *-ador*, *-adora*, that indicate the person who usually performs the action as an agent. Most of the time, it is composed by a verb followed by the suffix. Some examples are: “acopiador”, “comentador”, “cuidador”, “educador”, “excavador”, “investigador”, “oteador”, “pensador”, “techador”, “zelador”, “aviadora”, “embajadora”, “presentadora” and “rejoneadora”.
- *-ano*, which denotes “relative to” or “belonging to”, such as “artesano” and “franciscano”.
- *-ero*, *-era*, that is linked to trade, profession, place. It could also be assigned to utensils and tools: “carpintero”, “confitero” and “camarera”.
- *-ólogo*, *-óloga*, from the Greek “λογος” (“logos”) and the Latin “lŏgus”, meaning “specialist”: “antropólogo”, “biólogo”, “sociólogo”, “ginecóloga”, “podóloga”, “politóloga” and “sinóloga”.

- *-tor*, *-tora*, *-triz*, that refers to the position, job or dignity, the duration of this work and/or the place where it is performed: “cantor”, “directora”, “escritora”, “locutora”, “repositora”, “actriz”.

Whereas in English it was common to find occupations that seemed old in both masculine and feminine, the case of Spanish differs in that regard. More in detail, the tendency was to have those occupations exclusively in masculine-related occupations. For most part, it was the result of fastText (5 examples) and just one for GloVe and word2vec. The complete list will be presented at the end of this paragraph, using the Spanish Diachronic Corpus database (CORDE) of the Royal Spanish Academy (RAE)[12]: “acopiador” (1803 [fastText]), “timbalero” (1768 [fastText]), “thaumaturgo” (1737 [fastText]), “zelador” (1427 [fastText]), “baqueano” (1770 [fastText]), “confitero” (1626 [GloVe]), “capuchino” (1654 [word2vec]).

In contrast, the feminine-related words refer to new vocabulary. Let’s take some random occupations, using the same proportion of the language models, and apply the CORDE to them. Only four words are available[13]: “aviadora” (1747 [fastText]), “abogada” (1326 [fastText & word2vec]), “embajadora” (1602 [GloVe]), “presidenta” (1638 [word2vec]). In the first case, it has nothing to do with aeroplanes, but to the one who “plans, arranges or prepares something” (Royal Spanish Academy, n.d., definition 2). As per “abogada”, it denotes someone or something that plays a role of an “intercessor or mediator” (Royal Spanish Academy, n.d., definition 2). Lastly, “embajadora” and “presidenta” refer to “the wife” of an ambassador or a president (Royal Spanish Academy, n.d., definitions 4 and 7, respectively).

Since most they do not actually indicate occupations or professions, the alternative is to make a consultation in Current Spanish Reference Corpus (CREA)[14]. Here, the results vary: “podóloga” (1997 [fastText]), “sinóloga” (1996 [fastText]), “ginecóloga” (1983 [fastText]), “aviadora” (1923 [fastText]), “abogada” (1992 [fastText & word2vec]), “embajadora” (1989 [GloVe]), “presidenta” (1994 [word2vec]).

One possible hypothesis is that women were late entrants to the labour force and that this is the reason why the professions are more “modern”. However, this does not explain why there is such an older selection of masculine-related words.

[12] Available at http://corpus.rae.es/cordenet.html. Last accessed September 2021.
**N. B.:** To choose these dates, the oldest available reference was used.
[13] Idem.
[14] Available at http://corpus.rae.es/creanet.html. Last accessed September 2021.
**N. B.:** To choose these dates, the oldest reference available was used.

## Extreme Feminine Professions in Spanish Statistics

| | fastText | GloVe | word2vec |
|---|---|---|---|
| *Asistente* | -0,0107 | -0,0246 | -0,0091 |
| *Costurera* | -0,4618 | -0,5247 | -0,4225 |
| *Bibliotecaria* | -0,5006 | -0,3468 | -0,4032 |
| *Archivista* | 0,0293 | 0,0732 | 0,0620 |
| *Educadora* | -0,6193 | -0,5541 | -0,5756 |
| *Veterinaria* | -0,0430 | -0,0243 | -0,0540 |
| *Jardinera* | -0,3344 | -0,4185 | -0,2778 |
| *Auxiliar* | 0,0653 | 0,0941 | 0,0834 |
| *Gestora* | -0,3234 | -0,3189 | -0,3033 |
| *Investigadora* | -0,4672 | -0,5045 | -0,5064 |
| *Low Bias Limit (fem)* | 0 — –0,0315 | 0 — –0,0565 | 0 — –0,0281 |
| *Medium Bias Limit (fem)* | –0,0316 — – 0,0693 | –0,0566 — – 0,1169 | –0,0282 — – 0,0625 |
| *High Bias Limit (fem)* | –0,0694 — – 0,1265 | –0,1170 — – 0,1989 | –0,0626 — – 0,1161 |
| *Extremely High Bias Limit (fem)* | Less than – 0,1266 | Less than – 0,1990 | Less than – 0,1162 |
| *Low Bias Limit (mas)* | 0 – 0,0359 | 0 — 0,0390 | 0 — 0,0299 |
| *Medium Bias Limit (mas)* | 0,0360 – 0,0751 | 0,0391 — 0,0849 | 0,0300 — 0,0631 |
| *High bias Limit (mas)* | 0,0752 – 0,1267 | 0,0850 — 0,1485 | 0,0632 — 0,1079 |
| *Extremely High Bias Limit (mas)* | More than 0,1268 | More than 0,1486 | More than 0,1080 |

| | fastText | GloVe | word2vec |
|---|---|---|---|
| *Asistente* | = | = | = |
| *Costurero* | -0,1467 | -0,2676 | -0,0549 |
| *Bibliotecario* | 0,1503 | 0,2561 | 0,2344 |
| *Archivista* | = | = | = |
| *Educador* | 0,3099 | 0,5030 | 0,3381 |
| *Veterinario* | 0,1407 | 0,2695 | 0,2466 |
| *Jardinero* | 0,2207 | 0,3501 | 0,2490 |
| *Auxiliar* | = | = | = |
| *Gestor* | 0,3278 | 0,4036 | 0,3211 |
| *Investigador* | 0,2690 | 0,4772 | 0,2354 |
| *Low Bias Limit (fem)* | 0 — –0,0315 | 0 — –0,0565 | 0 — –0,0281 |
| *Medium Bias Limit (fem)* | –0,0316 — – 0,0693 | –0,0566 — – 0,1169 | –0,0282 — – 0,0625 |
| *High Bias Limit (fem)* | –0,0694 — – 0,1265 | –0,1170 — – 0,1989 | –0,0626 — – 0,1161 |
| *Extremely High Bias Limit (fem)* | Less than – 0,1266 | Less than – 0,1990 | Less than – 0,1162 |
| *Low Bias Limit (mas)* | 0 – 0,0359 | 0 — 0,0390 | 0 — 0,0299 |
| *Medium Bias Limit (mas)* | 0,0360 – 0,0751 | 0,0391 — 0,0849 | 0,0300 — 0,0631 |
| *High bias Limit (mas)* | 0,0752 – 0,1267 | 0,0850 — 0,1485 | 0,0632 — 0,1079 |
| *Extremely High Bias Limit (mas)* | More than 0,1268 | More than 0,1486 | More than 0,1080 |

***Table 16.*** *Semantic orientation results for extreme feminine occupations and their masculine counterparts in Spanish by quartiles, using the words obtained from the National Statistics Institute of Spain.*

As previously mentioned, the information from **Table 16** contains a set of extreme feminine professions in Spanish statistics along with their masculine counterparts, made with the data from the National Statistics Institute (INE) of Spain and listed according to the job distribution of women. Its creation was done by taking the top female-related job and organising them according to their semantic orientation, as computed by three different language models (fastText, GloVe and word2vec) with the Spanish Billion Word Corpus as a base. The quartile classification explained in preceding sections was applied here as well.

Whilst English followed the pattern of biased professions obtained by Bolukbasi et al. (2016), the overlapping was reduced in the case of the Spanish language. On this occasion, only "bibliotecaria" remained present. However, there are some similarities with the experiments in the previous section as well.

Firstly, an irregular pattern was identified in terms of the assignment of semantic orientation, which was also present in English. Also, a total of 6 occupations achieved an extremely feminine-biased rating. For "veterinaria", there was a trend towards intermediate values, whereas in "asistente" a low bias tendency was identified. In contrast, "archivista" and "auxiliar" obtained male-related values, particularly the latter.

One fact to consider is that Spanish words have gender, which usually coincides with the person in question. For the masculine equivalents of feminine-related professions, it was shown the same conduct than in English, that is to say: they display a more constant behaviour and less variation than their feminine-related counterparts. A case that should be noted is the one of "costurera"-"costurero". Here, it is hard to determine if the high scores in both its masculine and feminine forms are determined by Grammar or by Semantics. Further research must be carried out in order to come to a definite conclusion.

Now, it is worth noting epicenes. These refer to terms whose gender is determined by the context, usually by the article that accompanies them. On their own, it is impossible to know whether they refer to a man or a woman. Nevertheless, as explained above, "archivista" and "auxiliar" had more masculine-related scores. Meanwhile, "asistente" exhibited a masculine tendency, albeit much more neutral than the other two.

When analysing the performance of the three studied language models, fastText, GloVe and word2vec practically mirror their results. The total number of high or extremely high feminine values in extreme feminine professions is the same as total number of high or extremely high masculine values in extreme masculine professions: 6 each. It means that it cannot be concluded that there is bias in this regard.

## Extreme Masculine Professions in Spanish Statistics

| | fastText | GloVe | word2vec |
|---|---|---|---|
| *Constructor* | 0,2977 | 0,3897 | 0,3263 |
| *Carpintero* | 0,2338 | 0,3380 | 0,4356 |
| *Ingeniero* | 0,2507 | 0,3821 | 0,2046 |
| *Pescador* | 0,1632 | 0,3195 | 0,3895 |
| *Zapatero* | 0,0451 | -0,0039 | 0,2386 |
| *Ebanista* | 0,1055 | 0,3600 | 0,2764 |
| *Reparador* | 0,2730 | 0,2429 | 0,1522 |
| *Agricultor* | 0,2292 | 0,3850 | 0,3021 |
| *Ganadero* | 0,1950 | 0,3352 | 0,2390 |
| *Informático* | 0,1250 | 0,3207 | 0,1318 |
| *Low Bias Limit (fem)* | 0 — –0,0315 | 0 — –0,0565 | 0 — –0,0281 |
| *Medium Bias Limit (fem)* | –0,0316 — – 0,0693 | –0,0566 — – 0,1169 | –0,0282 — – 0,0625 |
| *High Bias Limit (fem)* | –0,0694 — – 0,1265 | –0,1170 — – 0,1989 | –0,0626 — – 0,1161 |
| *Extremely High Bias Limit (fem)* | Less than – 0,1266 | Less than – 0,1990 | Less than – 0,1162 |
| *Low Bias Limit (mas)* | 0 – 0,0359 | 0 — 0,0390 | 0 — 0,0299 |
| *Medium Bias Limit (mas)* | 0,0360 – 0,0751 | 0,0391 — 0,0849 | 0,0300 — 0,0631 |
| *High bias Limit (mas)* | 0,0752 – 0,1267 | 0,0850 — 0,1485 | 0,0632 — 0,1079 |
| *Extremely High Bias Limit (mas)* | More than 0,1268 | More than 0,1486 | More than 0,1080 |

| | fastText | GloVe | word2vec |
|---|---|---|---|
| *Constructora* | -0,1767 | -0,2271 | -0,2009 |
| *Carpintera* | -0,0408 | -0,2147 | -0,0222 |
| *Ingeniera* | -0,5628 | -0,4688 | -0,5941 |
| *Pescadora* | -0,2775 | -0,1649 | -0,0388 |
| *Zapatera* | -0,1926 | -0,1177 | -0,3429 |
| *Ebanista* | = | = | = |
| *Reparadora* | -0,2603 | -0,2674 | -0,2035 |
| *Agricultora* | -0,3903 | -0,2049 | -0,2589 |
| *Ganadera* | 0,0058 | -0,0588 | -0,0322 |
| *Informática* | -0,0322 | -0,0480 | -0,0695 |
| *Low Bias Limit (fem)* | 0 — –0,0315 | 0 — –0,0565 | 0 — –0,0281 |
| *Medium Bias Limit (fem)* | –0,0316 — – 0,0693 | –0,0566 — – 0,1169 | –0,0282 — – 0,0625 |
| *High Bias Limit (fem)* | –0,0694 — – 0,1265 | –0,1170 — – 0,1989 | –0,0626 — – 0,1161 |
| *Extremely High Bias Limit (fem)* | Less than – 0,1266 | Less than – 0,1990 | Less than – 0,1162 |
| *Low Bias Limit (mas)* | 0 – 0,0359 | 0 — 0,0390 | 0 — 0,0299 |
| *Medium Bias Limit (mas)* | 0,0360 – 0,0751 | 0,0391 — 0,0849 | 0,0300 — 0,0631 |
| *High bias Limit (mas)* | 0,0752 – 0,1267 | 0,0850 — 0,1485 | 0,0632 — 0,1079 |
| *Extremely High Bias Limit (mas)* | More than 0,1268 | More than 0,1486 | More than 0,1080 |

***Table 17.*** *Semantic orientation results for extreme masculine occupations and their feminine counterparts in Spanish by quartiles, using the words obtained from the National Statistics Institute of Spain.*

The same way in happened in the extreme masculine occupations in English (**Table 8**), in **Table 17** there is a clear tendency of job nouns to be highly regarded as extremely biased towards masculine values. It is not totally unanimous, since “informático” and, particularly “zapatero” have a more tempered scores that lean towards medium and low bias, respectively.

Once again, the extreme feminine occupation list demonstrates an irregular pattern in the assignment of gender bias quartiles, which is now a feature of this category. In total, 5 of the words achieved a high feminine-related score. For the cases of “ganadera” and “informática”, their values are more neutral. This is also true for two terms: “carpintera” and “pescadora”, but to a lesser extent. The only epicene in this list, namely “ebanista”, revolves around a male-related score.

In regards to the behaviour of the language models, there are differences according to the type of occupations to which it refers. In terms of the masculine-related jobs, word2vec allocates all words as extremely high biased. Next, comes GloVe (9) and fastText (8). There is not a single reference to female values in this group.

On the opposite side, the feminine counterparts reveal less intensity in the bias of the results. Here, GloVe (7) and word2vec (7) stand out from fastText (6). Plus, all three models generate one high or extremely high masculine values.

## Gender-Neutral Professions in Spanish Statistic

| | fastText | GloVe | word2vec |
|---|---|---|---|
| *Financiero* | 0,0274 | 0,1428 | 0,0009 |
| *Comerciante* | 0,1735 | 0,2499 | 0,2363 |
| *Científico* | 0,1748 | 0,3075 | 0,3699 |
| *Farmaceuta* | -0,2066 | -0,0077 | -0,1342 |
| *Editor* | 0,1756 | 0,2601 | 0,1014 |
| *Presentador* | 0,1428 | 0,3067 | 0,0303 |
| *Publicista* | -0,0344 | 0,0693 | -0,0425 |
| *Funcionario* | 0,1666 | 0,2211 | 0,1349 |
| *Artista* | -0,0417 | -0,0458 | -0,0101 |
| *Low Bias Limit (fem)* | 0 — –0,0315 | 0 — –0,0565 | 0 — –0,0281 |
| *Medium Bias Limit (fem)* | –0,0316 — –0,0693 | –0,0566 — –0,1169 | –0,0282 — –0,0625 |
| *High Bias Limit (fem)* | –0,0694 — –0,1265 | –0,1170 — –0,1989 | –0,0626 — –0,1161 |
| *Extremely High Bias Limit (fem)* | Less than –0,1266 | Less than –0,1990 | Less than –0,1162 |
| *Low Bias Limit (mas)* | 0 – 0,0359 | 0 — 0,0390 | 0 — 0,0299 |
| *Medium Bias Limit (mas)* | 0,0360 – 0,0751 | 0,0391 — 0,0849 | 0,0300 — 0,0631 |
| *High bias Limit (mas)* | 0,0752 – 0,1267 | 0,0850 — 0,1485 | 0,0632 — 0,1079 |
| *Extremely High Bias Limit (mas)* | More than 0,1268 | More than 0,1486 | More than 0,1080 |

| | fastText | GloVe | word2vec |
|---|---|---|---|
| *Financiera* | -0,0837 | -0,1145 | -0,1051 |
| *Comerciante* | = | = | = |
| *Científica* | -0,0414 | -0,1066 | -0,0198 |
| *Farmaceuta* | = | = | = |
| *Editora* | -0,5208 | -0,5945 | -0,6043 |
| *Presentadora* | -0,6383 | -0,7147 | -0,7037 |
| *Publicista* | = | = | = |
| *Funcionaria* | -0,5811 | -0,6330 | -0,6183 |
| *Artista* | = | = | = |
| *Low Bias Limit (fem)* | 0 — –0,0315 | 0 — –0,0565 | 0 — –0,0281 |
| *Medium Bias Limit (fem)* | –0,0316 — –0,0693 | –0,0566 — –0,1169 | –0,0282 — –0,0625 |
| *High Bias Limit (fem)* | –0,0694 — –0,1265 | –0,1170 — –0,1989 | –0,0626 — –0,1161 |
| *Extremely High Bias Limit (fem)* | Less than –0,1266 | Less than –0,1990 | Less than –0,1162 |
| *Low Bias Limit (mas)* | 0 – 0,0359 | 0 — 0,0390 | 0 — 0,0299 |
| *Medium Bias Limit (mas)* | 0,0360 – 0,0751 | 0,0391 — 0,0849 | 0,0300 — 0,0631 |
| *High bias Limit (mas)* | 0,0752 – 0,1267 | 0,0850 — 0,1485 | 0,0632 — 0,1079 |
| *Extremely High Bias Limit (mas)* | More than 0,1268 | More than 0,1486 | More than 0,1080 |

***Table 18.*** *Semantic orientation results for gender-neutral occupations in Spanish by quartiles, using the words obtained from the National Statistics Institute of Spain.*

The information from **Table 18** contains a set of gender-neutral terms in Spanish, extracted from the data from National Statistics Institute (INE) of Spain. Occupations were put in both masculine and feminine (when available), as to compare the changes that words experience in either gender.

There still the tendency that gives masculine-related occupations higher scores in terms of gender bias. So, for instance, "comerciante" (an epicene), "científico", "editor", "presentador" and "funcionario" get high values in this regard. In comparison, only "editora", "presentadora" and "funcionaria" achieve the same numbers in that regard.

In terms of medium-size bias, female-related jobs get more examples than their masculine counterparts. The only masculine-related term is "financiero", which is joined by "financiera" and "científica". However, 3 out of 4 epicenes get feminine scores: "publicista", "artista" and "farmaceuta".

In regards to the language models, GloVe gets the highest number of masculine-related results (6), whilst fastText and word2vec are equal in the number of female results (5 each).

| | **Extremely Masculine Occupations** | | | **Feminine Counterparts** | | |
|---|---|---|---|---|---|---|
| | **fastText** | **GloVe** | **word2vec** | **fastText** | **GloVe** | **word2vec** |
| **Total number of high or extremely high feminine values** | **0** | **0** | **0** | **6** | **7** | **7** |
| *Total number of high or extremely high masculine values* | *8* | *9* | *10* | *1* | *1* | *1* |
| | **Masculine Counterparts** | | | **Extremely Feminine Occupations** | | |
| | **fastText** | **GloVe** | **word2vec** | **fastText** | **GloVe** | **word2vec** |
| **Total number of high or extremely high feminine values** | **1** | **1** | **0** | **6** | **6** | **6** |
| *Total number of high or extremely high masculine values* | *6* | *6* | *6* | *0* | *1* | *1* |
| | **Gender-Neutral Occupations in Masculine** | | | **Gender-Neutral Occupations in Feminine** | | |
| | **fastText** | **GloVe** | **word2vec** | **fastText** | **GloVe** | **word2vec** |
| **Total number of high or extremely high feminine values** | **1** | **0** | **1** | **5** | **3** | **5** |
| *Total number of high or extremely high masculine values* | *5* | *6* | *4* | *1* | *1* | *1* |

***Table 19.*** *Total number of high or extremely high values per kind of occupation and different language models in Spanish.*

**Table 19** summarises the total number of high or extremely high values obtained for masculine- or feminine-related occupations in different language models. As seen there, the classification per number of biased results in masculine-related occupations is: word2vec (16), GloVe (12) and fastText (6). Meanwhile, the ranking is the same with most biased results in feminine-related occupations is: word2vec (18), GloVe (13) and fastText (11).

Unlike English, the results overlap amongst the different language models. This is obvious, since there is just a single corpus per language model. However, a very interesting finding lies in this table: the extreme feminine related job list is the only one that appears to provide equal distribution of high and extremely high values for the Spanish professions.

Then, comes the gender-neutral occupation list. It seems to display a similar behaviour than English, in the sense that it alternates the language models in terms of bias: if fastText looks to lean more towards feminine values, GloVe seems to do the same in masculine values, whilst word2vec goes back to feminine. Once more, as previous stated in page 51, there is not enough evidence in these experiments to conclude that there is gender bias.

Apparently, the extreme masculine professions are the more intense out of the three groups. This is a phenomenon that was already evident in the previous section. But this not only true for masculine-related occupations, since their feminine-related counterparts also show a pretty strong bias on their behalf. This is an exceptionality that needs to be addressed in future research.

Now, it is the time to analyse the way the semantic orientation would change the order of the occupation lists extracted from the National Statistics Institute of Spain. These are available in **Table 20**. Their calculation was made in the same way as before: the information of the extremely high and high bias was counted for each of the occupations available in the various language models studied, and then added.

The lists are made like this:

- Extreme *woman* occupations (5 jobs): 1 original feminine ("costurera"), 3 original neutral ("farmaceuta", "financiero", "presentador") and 1 original masculine ("zapatero").
- Extreme *man* occupations (9 jobs): 2 original feminine ("veterinaria", "auxiliar"), 2 original neutral ("científico", "comerciante") and 5 original masculine ("carpintero", "ebanista", "ganadero", "informático", "pescador").
- Gender-neutral occupations (15 jobs): 7 original feminine ("archivista", "asistente", "bibliotecaria", "educadora", "jardinera", "gestora", "investigadora"), 4 original

neutral (“artista”, “editor”, “funcionario”, “publicista”) and 4 original masculine (“agricultor”, “constructor”, “ingeniero”, “reparador”).

In English, the gender-neutral category lost most of its elements to the masculine classification. In Spanish, this is the one that gets most of the words, especially those from feminine origin. Masculine-related terms also become neutral. In the end, more than half of the jobs become of this kind.

In conclusion, by seeing these results, one could be induced to think that Spanish looks like a more gender-neutral language than English. But, given the nature of English, which lacks a gender mark, this assertion is not really fact-based.

It is true, though, that in contrast with the previous exercise, the original lists and the new ones differ in great manner. This is to be expected, since the composition changed so much.

| **Occupations’ list according to their statistical distribution** | | |
|---|---|---|
| Extreme *woman* occupations | Extreme *man* occupations | Neutral occupations[15] |
| asistente | constructor | financiero |
| costurera | carpintero | comerciante |
| bibliotecaria | ingeniero | científico |
| archivista | pescador | farmaceuta |
| educadora | zapatero | editor |
| veterinaria | ebanista | presentador |
| jardinera | reparador | publicista |
| auxiliar | agricultor | funcionario |
| gestora | ganadero | artista |
| investigadora | informático | |
| **Occupations’ list according to their semantic orientation distribution** | | |
| Extreme *woman* occupations | Extreme *man* occupations | Neutral occupations |
| costurera | veterinaria | asistente |
| zapatera | comerciante | archivista |
| farmaceuta | científico | publicista |
| financiera | científico | artista |
| presentadora | ganadero | bibliotecario |
| | auxiliar | educador |
| | carpintero | jardinero |
| | pescador | gestor |
| | informático | investigador |
| | | constructor |
| | | ingeniero |
| | | reparador |
| | | agricultor |
| | | editor |
| | | funcionario |

***Table 20.*** *Comparison between labour statistics and semantic orientation distributions in Spanish.*

[15] Idem. **N. B.:** There were not enough jobs or occupations in Spain that were equally distributed, so only those 9 joined the experiments.

## 4.2 Job Characterisation According to Gender

In this section, the author will find and characterise the adjectives that go along the way with the occupation nouns, both in real life and in the static models, as to add more semantic insights to the findings made so far. This experiment will not only concern the male- and female-related nouns that have been mentioned until now, but also incorporate scores for neutral words, as a way to recognise the differences with more extreme jobs. It was thought to be the only way to truly weigh the results without biasing them at the same time. That way, there could be a real comparison amongst the system outputs, evaluating what changes when modifying the gendered terms. The final aim is to grasp a deeper comprehension of how English and Spanish operate in this regard through their examination in fastText, GloVe and word2vec.

Once the professions were defined and analysed, it was time to label the neighbouring adjectives attached to them, since they express grammatical features in the form of permanent or temporary attributes. The procedure to obtain each of them is described as follows: for every individual occupation noun, the possible answer was manually taken from the first 1,000 words of the automatically-extracted lists of words. Gonen & Goldberg (2019) had already suggested a similar approach to the issue during the course of their research. They found how the k-nearest neighbour of a word in a male-female semantic orientation space was a method to tackle gender bias. The idea was that the implicit gendered word attached to the original term would provide an extra set of perceptions that are really hard to detect when using a vector plot alone.

Also, as previously stated in this project, by examining embeddings and word lists, it is possible to estimate the strength of connection between neutral words and a social group. Garg et al. (2018) had already referred to this link, finding that "a natural metric for the embedding bias is the average distance for women minus the average distance for men. If this value is negative, then the embedding more closely associates the ocscupation with men".

These researchers associated the dynamics of the embedding with the quantifiable changes in US society –e.g., demographic and occupation shifts, which served as a guiding light for this investigation. They concluded that the relation between embedding bias score and "reality", as calculated by occupation participation, is consistent over time. Besides, the occupations that possess a nearly 50-50 split in gender participation have a small embedding bias (Garg et al., 2018).

Garg et al. (2018) proposed to make comparative statements to study how the description of women through adjectives evolve over time. They argued that an application for this work could determine how various narratives and portrayals of women developed and competed as the years passed. This approach seems to be much productive than the analogy analysis which is often used to expose how strongly human biases are encoded in language models.

As previously stated, two main lexicographic resources where employed in order to avoid any lack of accuracy in the classification of adjectives. Those two are the Essential British Dictionary from Cambridge University Press and the Spanish Language Dictionary from the Royal Spanish Academy (RAE).

Even though the original impression was to automatically extract the adjectives using the Natural Language Toolkit, often known as NLTK[16], the practice showed that it was not precise enough. Whilst revising the results provided by that system versus the cross-checking with dictionaries proved that the approach had to be reformulated.

Next, **Table 21** will provide the information regarding the English experiments, whilst **Table 22** shows the obtained results for Spanish.

[16] NLTK is a free, open source, community-driven platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenisation, stemming, tagging, parsing, and semantic reasoning.

Its documentation describes it as "suitable for linguists, engineers, students, educators, researchers, and industry users alike". NLTK is available for Windows, Mac OS X, and Linux.

Further information is available at: https://www.nltk.org/

| | Embedding professions | | | | Statistical professions | | Embedding Professions Ending in –Man, –Woman | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Extreme Masculine Professions | | Extreme Feminine Professions | | | | –Man Embedding Professions | | | –Woman Embedding Professions | | |
| | **Counterparts** | **Masculine** | **Feminine** | **Counterparts** | **Feminine** | **Masculine** | **–woman** | **–man** | **–person** | **–woman** | **–man** | **–person** |
| *Total number of adjectives:* | N/A | 87 | 122 | 60 | 65 | 86 | 68 | 131 | 20 | 68 | 67 | 72 |
| *Number of most mentioned adjectives:* | N/A | 6 | 10 | 4 | 7 | 4 | 2 | 4 | N/A | 11 | 9 | 2 |
| *Most common adjectives areas and total number of mentions:* | N/A | sports (20)<br>defence (13)<br>other (5) | religion (32)<br>personal features (19)<br>arts (5)<br>other (5) | religion (28) | medicine (39)<br>other occupations (12) | craftsmanship (17)<br>other (12) | sports (13) | sports (16)<br>defence (8) | N/A | politics (49)<br>personal features (15) | politics (45)<br>personal features (12)<br>money (8) | politics (13) |
| *Most common adjectives and total number of mentions:* | N/A | defensive 8<br>preseason 8<br>postseason 6<br>professional 6<br>best 5<br>offensive 5 | soulful 10<br>sultry 9<br>benedictine 7<br>carmelite 7<br>cistercian 7<br>monastic 6<br>augustinian 5<br>best 5<br>emerita 5<br>oscar-winning 5 | benedictine 8<br>distinguished 8<br>cistercian 6<br>monastic 6 | medical 13<br>nutritional 8<br>naturopathic 7<br>pediatric 7<br>paramedical 6<br>geriatric 5<br>paraprofessional 5 | artisan 12<br>itinerant 6<br>off-duty 6<br>skilled 5 | unseeded 7<br>top-seeded 6 | defensive 8<br>undrafted 6<br>preseason 5<br>unheralded 5 | N/A | mayoral 8<br>democratic 7<br>republican 7<br>congressional 6<br>gubernatorial 6<br>incumbent 6<br>née 5<br>non-executive 5<br>self-made 5<br>supervisory 5<br>well-connected 5 | republican 10<br>congressional 8<br>mayoral 8<br>wealthy 8<br>former 7<br>incumbent 7<br>gubernatorial 6<br>non-executive 5<br>prominent 5 | ex-officio 8<br>elect 5 |

***Table 21.*** *Summary of most common features in k-nearest adjective neighbours for feminine, masculine and gender-neutral professions in English.*

### 4.2.1. Adjectives in English

To build **Table 21**, there was the extraction for the different subsections:

- For the Embedding Professions: All the feminine-related and masculine-related professions were listed. Later on, the repeated terms were calculated and the number of repetitions each of them had was counted. Once that sum was done, the duplicates were removed and all the remaining professions of the list were counted individually. Then, the occupations that had four or more recurrences were fed into the system, in order to extract the k-nearest neighbours in each one of the language models (fastText, GloVe, word2vec) and corpora (original versions, Wikipedia, Gigawords, Wikipedia + Gigawords). Next, the adjectives for each subgroup of were counted. It is important to highlight they were summed separately. E.g., for the Extreme Feminine Professions, there was a collection of just feminine words (alumna, actress, chanteuse… etc.), and also another one for their masculine counterparts (alumnus, actor, chanteur… etc.). In the case of the Extreme Masculine Professions, there were no opposite-gender alternatives for "captain", "player", "actor" and "quarterback", and that is why the N/A appears.
- For the Statistical Professions: A random sample of four statistical occupations was selected. Since this project is just an exploratory study, it is out of its scope to include the comprehensive results for all the 30 extracted professions. For female purposes, "nutritionist", "nurse", "librarian" and "veterinarian" were utilised, whilst "mason", "firefighter", "carpenter" and "constructor" were used for masculine.
- For the Embedding Professions Ending in -Man and -Woman: During the process of choosing the word embedding professions, it was noticed that the system also provided many words ending in -man or -woman. So, the author decided to take advantage of that information in order to complement his findings in this section. Here, the N/A means that there were no repeated words for this category, so it was impossible to calculate a value.

The first finding in this section is that it looks like masculine-related occupations tend to generate more k-nearest neighbour as adjectives. The only exception is the one of -person-ending jobs in the -woman ending word embedding professions. Having said that, it seems like the highest quantity of most-mentioned adjectives belong to the feminine-related jobs. This pattern was already known, since it was evident in experiments regarding occupation nouns.

Regarding the topics of each classification, the extreme masculine embedding professions provide results in sports and defence, even though the adjectives from the latter could apply to the former as well. On the other hand, the extreme feminine embedding professions lean towards religion and with a very high number of most common adjectives, which repeat the prejudice that considers women as more pious than men. The majority of the most-common adjectives of this latest tag belong to religious orders, which is a think that looks from the past, as stated in the last section. It is also worth noticing the high weight of personal features in women-related results, like soulful or sultry.

The statistical professions, on their behalf, put a focus on medicine and other professions for feminine-related jobs, whilst their counterparts get a bit behind with topics linked to craftsmanship. This categorisation suggests that women are the people who care for others, and that is why one can find terms like “medical”, “nutritional” o “pediatric”. Men, do not seem to have the same worries, since they are assigned work such as “itinerant” or “off-duty”.

The same way it happened in the extreme occupations, the -man and -woman embedding occupations both offer the most mentioned adjectives in their respective categories. In terms of total quantities, it is the highest for men in their own classification, and the same is true for the -person-related set in the -woman-ending section. An interesting fact is that -man-ending has an emphasis on sports, whilst -woman-ending stresses the politics values. The most-linked sport to *her* is tennis, whilst the outcome for *him* is inconclusive. The -man embedding professions provides a total of 20 results for word -person-ending, but none of them repeats itself even twice.

Only in the extreme feminine embedding professions and in the -woman embedding professions one can find repeated for masculine and feminine. As mentioned before, in the former case it refers to religious orders, whilst in the latter it relates to politic terms. An insight that is worth noticing is how women are linked to the Democratic Party and men with the Republican. It also proves how US-centred the corpora are, since they only underscore American politics. In the case of women, they are described as “self-made” or “well-connected”, whereas the men words are linked to power derived from money: “wealthy”, “prominent”.

| | Embedding professions | | | | Statistical Professions | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Extreme Masculine Professions | | Extreme Feminine Professions | | Extreme Masculine Professions | | Extreme Feminine Professions | | Epicenes | |
| | **Counterparts** | **Masculine** | **Feminine** | **Counterparts** | **Counterparts** | **Masculine** | **Feminine** | **Counterparts** | **Feminine** | **Masculine** |
| *Total number of adjectives:* | 37 | 28 | 41 | 47 | 7 | 8 | 22 | 25 | 12 | 11 |
| *Number of most mentioned adjectives:* | 6 | 6 | 14 | 16 | 5 | 4 | 5 | 4 | 2 | 3 |
| *Most common adjectives areas and total number of mentions:* | intellectualism (7)<br>personal features (7) | personal features (12)<br>intellectualism (8) | entertainment (19)<br>employment (12)<br>education (8)<br>literature (7)<br>entrepreneurship (5)<br>personal features (5) | employment (18)<br>entertainment (15)<br>literature (9)<br>money & enterprise (9) | money & enterprise (11)<br>employment (2) | money & enterprise (6)<br>employment (5) | personal features (8)<br>literature (5) | literature (7)<br>education (2) | employment (5) | creativity (4)<br>employment (3) |
| *Most common adjectives and total number of mentions:* | anciana 3<br>estudiosa 3<br>admiradora 2<br>aventajada 2<br>feminista 2<br>librepensadora 2 | estudioso 5<br>aventajado 3<br>continuador 3<br>discapacitado 3<br>humanista 3<br>seguidor 3 | educadora 8<br>comunicadora 6<br>fundadora 5<br>nacida 5<br>conductora 4<br>narradora 4<br>coordinadora 3<br>delegada 3<br>investigadora 3<br>penalista 3<br>porno 3<br>reportera 3<br>televisivo 3<br>traductora 3 | asesor 4<br>cinematográfico 3<br>comunicador 3<br>coordinador 3<br>delegado 3<br>investigador 3<br>literario 3<br>millonario 3<br>multimillonario 3<br>penalista 3<br>propietario 3<br>radiofónico 3<br>reportero 3<br>teatral 3<br>televisivo 3<br>traductor 3 | concesionaria 3<br>constructora 3<br>constructoras 3<br>constructor 2<br>ferrovial 2 | constructor 3<br>constructora 3<br>propietario 3<br>constructores 2 | huérfana 3<br>humilde 3<br>traductora 3<br>bibliófila 2<br>soltera 2 | traductor 3<br>bibliófilo 2<br>bibliográfico 2<br>educador 2 | asesor 3<br>nombrado 2 | constructor 3<br>ilustrador 2<br>proyectista 2 |

***Table 22.*** *Summary of most common features in k-nearest adjective neighbours for feminine, masculine and gender-neutral professions in Spanish.*

#### 4.2.2. Adjectives in Spanish

The process to create **Table 22** follow the same instructions that were used for **Table 21**. The main difference was that instead of employing -man and -woman ending words, almost all were using the male and female equivalents in Spanish. The exception were the epicenes, which require context to be assessed in terms of gender, so they seem a nice way to compare their behaviour in regards to the gender-marked occupations.

- Embedding Professions:
  - Female: empresaria/o, abogada/o, actriz/actor, director/a, escritor/a, presentador/a
  - Male: discípulo/a, cuidador/a, pensador/a
- Statistical Professions:
  - Female: costurera/o, bibliotecaria/o
  - Male: constructor/a
- Epicenes:
  - Feminine: asistente/a
  - Masculine: ebanista

Unlike English, Spanish has the same number of total number of adjectives (119) for masculine and feminine occupations, but their distribution changes according to the kind of category. Feminine-related occupations prevail in extreme masculine professions (embedding), as well as in the statistical professions: extreme feminine and epicenes. Masculine-related jobs rank better in extreme feminine professions (for both embedding and statistical lists) and the extreme masculine professions from the statistics.

In short: the behaviour is very erratic, so it is hard to predict a pattern.

Then, the number of most mentioned adjectives seems to go hand to hand: 32 for female-related occupations and 33 for masculine-related. They even reached the same distribution in extreme masculine professions from the embeddings, and the results replicated exactly in statistical lists, in the extreme masculine and extreme feminine occupations.

Regarding the topics of each classification, they also repeat themselves. For instance, "employment" is a topic that repeats again and again, without making any kind of discrimination: feminine, masculine or epicene, embedding or statistical. The opposite is true for entertainment, which seems to be linked to extreme feminine professions in word embeddings, whilst “education” and “literature” are themes linked to all kinds of extreme feminine professions. Finally, nearly 2 out of 3 mentions of “money & enterprise” are related to extreme masculine professions in statistics.

The most common adjectives seem to mirror themselves in the different categories. For instance, in the extreme masculine embedding professions, there are two exact coincidences (“estudioso/a” and “aventajado/a”), a pair of synonyms (“admiradora”, “seguidor”) and an idea that women look after “ancianas” and men “discapacitados”. The most interesting aspects are the differences, though. There is an opposition of “humanista” and “feminista”, as if women were not humans or as if feminism was not another perspective from the humanism. Also, the system antagonises “librepensadora” with “continuador”, which is an interesting conflict to make.

On the other hand, the same exercise can be done in the extreme feminine embedding professions. There, half of the terms were repeated but, the semantical insights come from the diversity. For example, whilst a woman is the “fundadora” of a company, the man is a “propietario”, who also happens to be “millonario” or “multimillonario” even. Our lady works as an “educadora”, “conductora” or “narradora” and the gentleman as an “asesor” “cinematográfico”, “literario”, “radiofónico” or “teatral”. Lastly, our imaginary female subject probably works in “porno” … even though the chance for a man to do something “pornográfico” is 3 times less likely.

In the case of the extreme masculine statistical professions, something curious occurred: sure, there was the typical repetition of terms, but also the fact that “constructor” seemed to refer to a person, whilst “constructora” had a sense of an enterprise (“concesionaria”, “ferrovial”).

Lastly, the extreme feminine statistical occupations gave an emphasis on personal matters rather than keeping it professional. Terms like “huérfana”, “humilde” and “soltera” referred to woman, and this was the only instance where such a thing happened.

So, as to summarise the results of this section, it is true that Spanish apparently shows a more equal behaviour than English at first sight. However, when looking closely, one can still tell the way bias lies underneath, reflecting a reality of inequality towards women. This happens either by disregarding their full condition as humans, by echoing their lack of resources in regards to men or by placing aside their value as competent professionals with *ad hominem* fallacies.

# 5. Conclusions and Implications

This project explored three static word embeddings in terms of gender bias. Using the previous research of Bolukbasi et al. (2016), Caliskan et al. (2017) and, particularly, Garg et al. (2018) in English, their findings were applied to Spanish as well. The aim was to see if there were differences in those languages in GloVe, word2vec and fastText, and also between them. To do so, two linguistic data sets were created, using the information from the U.S. Labor Statistics and the National Statistics Institute of Spain for semantic orientation. Additionally, self-made lists of extreme masculine and feminine nouns and adjectives were extracted from three word embedding spaces. The research done in this report compared all of those, as to determine what varied in regards to different two groups and languages.

In terms of WE behaviour, it is hard to establish the way they work or to establish a correlation with the statistical information. As explained during the course of this project, the cases of jobs in the semantic orientation experiments for extreme professions only demonstrate one pattern: there is indeed a strong semantic orientation link in masculine-related words. This is not only true for masculine-related occupations, since their feminine-related counterparts also show a pretty strong bias on their behalf. This is something that needs to be addressed in future research.

Furthermore, whilst in English the extreme female occupations appeared to stay almost the same, most of the gender-neutral jobs seemed to transform into extreme masculine professions, whereas the bias in masculine terms remained unscathed. In Spanish, the majority of extreme female occupations became gender-neutral. At the same time, the extreme male occupations' inventory stayed constant in at least half of the cases, losing one to the female occupation list and four to the gender-neutral. Finally, the gender-neutral catalogue lost half of its words, but gained 11 from other sets.

There are more masculine-related professions than female-linked jobs. Even so, the proportion English-Spanish stays in a very similar proportion (masculine: 26 vs. 41 - feminine: 23 vs. 37).

In regards to biased results in English, the extreme feminine occupations, word2vec had the most prejudice with 8 out of 10 professions were considered as highly or extremely high. In extreme masculine occupations, fastText provided 9 out of 10 professions with the Wikipedia and the Wikipedia + Gigawords corpora. Also, GloVe in the Wikipedia configuration had the same score. For gender-neutral, GloVe had the same number with the Wikipedia and the Gigawords corpora: 8 professions.

Meanwhile, in Spanish the classification per number of biased results in masculine-related occupations was word2vec (16), GloVe (12) and fastText (6); whereas, the ranking is similar in feminine-related occupations: word2vec (18), GloVe (13) and fastText (11). Here, epicenes had more masculine-leaning scores.

Whereas in English it was common to find occupations that seemed old in both masculine and feminine, the case of Spanish differs in that regard. More in detail, the tendency was to have those occupations exclusively in masculine-related occupations. For most part, it was the result of fastText (5 examples) and just one for GloVe and word2vec. In contrast, female-related jobs seem to be freshly coined words, most of them from the 1980's onwards.

In terms of adjective characterisation, although masculine-related occupations in English tend to generate more total number of adjectives, it seems like the highest quantity of most-mentioned adjectives belong to the feminine-related jobs. On the other hand, Spanish has a mirroring total number of adjectives for masculine and feminine occupations (119). Also, the number of most mentioned adjectives seems to go hand to hand: 32 for female-related occupations and 33 for the masculine-related.

In English the extreme masculine embedding professions tend to provide sports and defence-related qualifiers, whilst extreme feminine embedding professions give preference to religion and politics, to a lesser extent. The statistical professions put a focus on medicine and other professions for feminine-related jobs, whilst their counterparts get a bit behind with topics linked to craftsmanship.

Then, in Spanish "employment" is a theme that repeats again and again, without discriminating amongst feminine, masculine or epicene; embedding or statistical. The opposite is true for "entertainment", which seems to be linked to extreme feminine professions in word embeddings. Finally, nearly 2 out of 3 mentions of "money & enterprise" are related to extreme masculine professions in statistics.

As an end note, further research should be done to make comparative statements about how the description of women through adjectives and word embedding spaces evolve over time. As Garg et al. (2018) argued, these investigations could explain the way narratives and portrayals of women develop and compete as the years pass, in English, Spanish and also other languages.

# 6. References


Amrami, A., & Goldberg, Y. (2018). Word sense induction with neural biLM and symmetric patterns. *arXiv preprint arXiv:1808.08518*.

Amrami, A., & Goldberg, Y. (2019). Towards better substitution-based word sense induction. *arXiv preprint arXiv:1905.12598*.

Aristotle, ., & Lord, C. (2013). *Aristotle's Politics*.

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (August 15, 2003). A Neural Probabilistic Language Model. *Journal of Machine Learning Research, 3,* 6, 1137-1155.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135-146.

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in neural information processing systems* (pp. 4349-4357).

Brunet, M. E., Alkalay-Houlihan, C., Anderson, A., & Zemel, R. (2019, May). Understanding the origins of bias in word embeddings. In *International Conference on Machine Learning* (pp. 803-811).

Bühler, K., & Marías, A. J. (1950). *Teoría del lenguaje*. Madrid: Revista de Occidente.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183-186.

Camacho-Collados, J., & Pilehvar, M. T. (May 10, 2018). From word to sense embeddings: A survey on vector representations of meaning. *Arxiv*.

Chada, R. (2019). Gendered Pronoun Resolution using BERT and an extractive question answering formulation. *arXiv preprint arXiv:1906.03695*.

Chen, D., & Manning, C. D. (2014, October). A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 740-750).

Crawford, K. (2017, December). The trouble with bias. In *Conference on Neural Information Processing Systems, invited speaker*.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fischer, O. (2019). *Analogy in Language and Linguistics*. Oxford University Press.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (March 12, 2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences of the United States of America, 115,* 16, 3635.

Gonen, H., & Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862*.

Gonen, H., & Webster, K. (2020). Automatically identifying gender issues in machine translation using perturbations. *arXiv preprint arXiv:2004.14065*.

Harper, D. (n.d.) -ess. In *Online Etymology Dictionary*. Retrieved on 15 June, 2021 from https://www.etymonline.com/word/-ess

Harris, Z. S. (1954). Distributional structure. *WORD*, 10(2-3):146–162.

Hatzivassiloglou, V., & McKeown, K. (1997, July). Predicting the semantic orientation of adjectives. In 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics (pp. 174-181).

Herbelot, A., Von Redecker, E., & Müller, J. (2012, April). Distributional techniques for philosophical enquiry. In *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (pp. 45-54).

Huang, L., Sun, C., Qiu, X., & Huang, X. (2019). GlossBERT: BERT for word sense disambiguation with gloss knowledge. *arXiv preprint arXiv:1908.07245*.

Joshi, M., Levy, O., Weld, D. S., & Zettlemoyer, L. (2019). BERT for coreference resolution: Baselines and analysis. *arXiv preprint arXiv:1908.09091*.

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Kurita, K., Vyas, N., Pareek, A., Black, A. W., & Tsvetkov, Y. (2019). Measuring bias in contextualized word representations. *arXiv preprint arXiv:1906.07337*.

Levy, O., & Goldberg, Y. (2014, June). Linguistic regularities in sparse and explicit word representations. In *Proceedings of the eighteenth conference on computational natural language learning* (pp. 171-180).

Liu, B. (2019). Anonymized BERT: An Augmentation Approach to the Gendered Pronoun Resolution Challenge. *arXiv preprint arXiv:1905.01780*.

Loureiro, D., & Jorge, A. (2019). Language modelling makes sense: Propagating representations through wordnet for full-coverage word sense disambiguation. *arXiv preprint arXiv:1906.10007*.

Mikolov, Tomas, Chen, Kai, Corrado, Greg, & Dean, Jeffrey. (2013a). *Efficient Estimation of Word Representations in Vector Space*.

Mikolov, Tomas, Le, Quoc V., & Sutskever, Ilya. (2013b). *Exploiting Similarities among Languages for Machine Translation*.

Mikolov, Tomas, Sutskever, Ilya, Chen, Kai, Corrado, Greg, & Dean, Jeffrey. (2013c). *Distributed Representations of Words and Phrases and their Compositionality*.

Neelakantan, A., Shankar, J., Passos, A., & McCallum, A. (2015). *Efficient Non-parametric Estimation of Multiple Embeddings per Word in Vector Space*.

New International Version. (2011). https://www.biblegateway.com/passage/?search=Genesis+1&version=NIV

Nissim, M., van Noord, R., & van der Goot, R. (2020). Fair is better than sensational: Man is to doctor as woman is to doctor.

Orkphol, K., & Yang, W. (2019). Word sense disambiguation using cosine similarity collaborates with Word2vec and WordNet. *Future Internet*, *11*(5), 114.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532– 1543, Doha, Qatar. Association for Computational Linguistics.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Prates, M. O., Avelar, P. H., & Lamb, L. C. (2019). Assessing gender bias in machine translation: a case study with Google Translate. *Neural Computing and Applications*, 1-19.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Rahutomo, F., Kitasuka, T., & Aritsugi, M. (2012, October). Semantic cosine similarity. In The 7th International Student Conference on Advanced Science and Technology ICAST (Vol. 4, No. 1, p. 1).

Royal Spanish Academy. (n.d.). abogada. In *Dictionary of the Spanish Language*. Retrieved from: https://dle.rae.es/abogado?m=form

Royal Spanish Academy. (n.d.). aviadora. In *Dictionary of the Spanish Language*. Retrieved from: https://dle.rae.es/aviador#4ZCbk71

Royal Spanish Academy. (n.d.). embajadora. In *Dictionary of the Spanish Language*. Retrieved from: https://dle.rae.es/embajador?m=form

Royal Spanish Academy. (n.d.). presidenta. In *Dictionary of the Spanish Language*. Retrieved from: https://dle.rae.es/presidenta

Salton, G., Wong, A., & Yang, C. S. (November 01, 1975). A vector space model for automatic indexing. *Communications of the Acm, 18,* 11, 613-620.

Saunders, D., Sallis, R., & Byrne, B. (2020). Neural Machine Translation Doesn't Translate Gender Coreference Right Unless You Make It. *arXiv preprint arXiv:2010.05332*.

Saussure, F. ., & Alonso, A. (1945). *Curso de Lingüística General*. Buenos Aires: Editorial Losada.

Schütze. H. (1998). Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

Seo, M., Kembhavi, A., Farhadi, A., & Hajishirzi, H. (2016). Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.

Shannon, C. E. (1951). Prediction and entropy of printed English. *The Bell System Technical Journal*, 30(1):50–64.

Sommaggio, P., Mazzocca, M., Gerola, A., & Ferro, F. (2004). Cognitive liberty. A first step towards a human neuro-rights declaration. *Annals of the New York Academy of Sciences*, *1013*(1), 221-228.

Stalin, I. V. (1977). *Obras escogidas*. Madrid: Emiliano Escolar.

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019, October). How to fine-tune bert for text classification?. In *China National Conference on Chinese Computational Linguistics* (pp. 194-206). Springer, Cham.

Trask, R. L. (1996). Adjective. In *A Dictionary of Grammatical Terms in Linguistics*. (pp. 7-8). London: Routledge.

Turney, P. D., & Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, *21*(4), 315-346.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Wang, Y., Huang, M., Zhu, X., & Zhao, L. (2016, November). Attention-based LSTM for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 606-615).

Wang, Z. (2019). MSnet: A BERT-based Network for Gendered Pronoun Resolution. *arXiv preprint arXiv:1908.00308*.

Wittgenstein, L., Anscombe, G. E. M., In Rhees, R., & Ramsey, I. T. (1953). Philosophical investigations. Oxford: Blackwell.

Yaghoobzadeh, Y., & Schütze, H. (2016). *Intrinsic Subspace Evaluation of Word Embedding Representations*.

Yang, K. C., Niven, T., Chou, T. H., & Kao, H. Y. (2019, August). Fill the GAP: Exploiting BERT for Pronoun Resolution. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing* (pp. 102-106).

Zhao, J., Wang, T., Yatskar, M., Cotterell, R., Ordonez, V., & Chang, K. W. (2019). Gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.03310*.

Zhao, J., Zhou, Y., Li, Z., Wang, W., & Chang, K. W. (2018). Learning gender-neutral word embeddings. *arXiv preprint arXiv:1809.01496*.

Zhelezniak, V., Shen, A., Busbridge, D., Savkov, A., & Hammerla, N. (2019). Correlations between word vector sets. *arXiv preprint arXiv:1910.02902*.

# 7. Appendices

## Semantic Orientation Quartile Distribution for Extreme Masculine Professions in U.S. Labour Statistics

| | **Extreme Masculine Professions in U.S. Labour Statistics** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | fastText | | | GloVe | | | | word2vec |
| | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| *Mason* | 0,2626 | 0,2933 | 0,2787 | 0,2797 | 0,1322 | 0,2036 | 0,2455 | 0,3117 |
| *Plumber* | 0,3320 | 0,2643 | 0,2975 | 0,1124 | 0,1667 | 0,2818 | 0,2517 | 0,2395 |
| *Mechanic* | 0,2279 | 0,4261 | 0,3082 | 0,2637 | 0,2101 | 0,4331 | 0,4077 | 0,3289 |
| *Roofer* | 0,2865 | 0,2174 | 0,3156 | 0,0886 | 0,1821 | 0,1478 | 0,1063 | 0,3223 |
| *Millwright* | 0,2468 | N/A | 0,2422 | -0,0550 | 0,1417 | N/A | 0,1100 | 0,0755 |
| *Firefighter* | 0,1878 | 0,2065 | 0,2703 | 0,1543 | 0,2674 | 0,3207 | 0,3322 | 0,2186 |
| *Carpenter* | 0,2677 | 0,3104 | 0,3423 | 0,3455 | 0,3159 | 0,2725 | 0,4058 | 0,3582 |
| *Pilot* | 0,1141 | 0,0870 | 0,2096 | 0,2064 | 0,2995 | 0,2895 | 0,3818 | 0,0708 |
| *Constructor* | 0,0186 | 0,0584 | 0,0950 | -0,0588 | 0,0192 | -0,0645 | 0,0334 | -0,0330 |
| *Repairer* | 0,2835 | 0,2164 | 0,2568 | 0,1159 | 0,2399 | 0,1118 | 0,0646 | 0,0948 |
| *Low Bias Limit (fem)* | 0 – –0.0287 | 0 – -0.0305 | 0 – -0.0303 | 0 – -0.0729 | 0 – -0.0502 | 0 – -0.0761 | 0 – -0.0764 | 0 – -0.0423 |
| *Medium Bias Limit (fem)* | -0.0288 – -0.0623 | -0.0306 – -0.0683 | -0.0304 – -0.0663 | -0.0730 – -0.1413 | -0.0503 – -0.1014 | -0.0762 – -0.1450 | -0.0765 – -0.1437 | -0.0424 – -0.0891 |
| *High Bias Limit (fem)* | -0.0624 – -0.1083 | -0.0684 – -0.1268 | -0.0664 – -0.1192 | -0.1414 – -0.2231 | -0.1015 – -0.1655 | -0.1451 – -0.2263 | -0.1438 – -0.2227 | -0.0892 – -0.1546 |
| *Extremely High Bias Limit (fem)* | Less than -0.1084 | Less than -0.1269 | Less than -0.1193 | Less than -0.2232 | Less than -0.1656 | Less than -0.2264 | Less than -0.2228 | Less than -0.1547 |
| *Low Bias Limit (mas)* | 0 – 0.0461 | 0 – 0.0502 | 0 – 0.0500 | 0 – 0.0344 | 0 – 0.0337 | 0 – 0.0347 | 0 – 0.0331 | 0 – 0.0395 |
| *Medium Bias Limit (mas)* | 0.0461 – 0.0954 | 0.0503 – 0.1016 | 0.0500 – 0.1021 | 0.0345 – 0.0783 | 0.0338 – 0.0739 | 0.0348 – 0.0795 | 0.0331 – 0.0756 | 0.0396 – 0.0866 |
| *High bias Limit (mas)* | 0.0955 – 0.1602 | 0.1017 – 0.1686 | 0.1022 – 0.1710 | 0.0784 – 0.1451 | 0.0740 – 0.1315 | 0.0796 – 0.1481 | 0.0757 – 0.1433 | 0.0867 – 0.1579 |
| *Extremely High Bias Limit (mas)* | More than 0.1603 | More than 0.1687 | More than 0.1711 | More than 0.1452 | More than 0.1316 | More than 0.1482 | More than 0.1434 | More than 0.1580 |

## Semantic Orientation Quartile Distribution for Extreme Feminine Professions in U.S. Labour Statistics

| | Extreme Feminine Professions in U.S. Labour Statistics | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | fastText | | | GloVe | | | | word2vec |
| | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| *Nutritionist* | -0.2831 | -0.3308 | -0.2529 | -0.3341 | -0.3051 | -0.3045 | -0.3378 | -0.3162 |
| *Hairdresser* | 0.1064 | -0.1763 | -0.0220 | -0.1509 | -0.0208 | -0.1646 | -0.0762 | -0.3548 |
| *Secretary* | -0.0716 | -0.0673 | -0.0379 | 0.0566 | -0.0807 | 0.0234 | 0.0321 | -0.1839 |
| *Nurse* | -0.3204 | -0.2796 | -0.2256 | -0.2450 | -0.1942 | -0.1612 | -0.1685 | -0.5054 |
| *Receptionist* | -0.0155 | -0.2426 | -0.1177 | -0.2731 | -0.1065 | -0.3315 | -0.1828 | -0.4995 |
| *Bookkeeper* | 0.0987 | -0.0670 | 0.1112 | -0.2203 | 0.0214 | -0.1229 | -0.1575 | -0.3307 |
| *Caretaker* | 0.0381 | 0.0628 | 0.0898 | 0.1944 | 0.0569 | 0.1352 | 0.1064 | 0.1816 |
| *Librarian* | -0.1959 | -0.2609 | -0.2298 | -0.2715 | -0.1394 | -0.2752 | -0.2586 | -0.4740 |
| *Veterinarian* | 0.0300 | -0.0752 | 0.0141 | -0.0911 | 0.0240 | 0.0195 | -0.0036 | -0.1031 |
| *Aide* | 0.1747 | 0.1588 | 0.2828 | 0.1152 | 0.1215 | 0.2605 | 0.3359 | -0.0090 |
| *Low Bias Limit (fem)* | 0 – –0.0287 | 0 – -0.0305 | 0 – -0.0303 | 0 – -0.0729 | 0 – -0.0502 | 0 – -0.0761 | 0 – -0.0764 | 0 – -0.0423 |
| *Medium Bias Limit (fem)* | -0.0288 – -0.0623 | -0.0306 – -0.0683 | -0.0304 – -0.0663 | -0.0730 – -0.1413 | -0.0503 – -0.1014 | -0.0762 – -0.1450 | -0.0765 – -0.1437 | -0.0424 – -0.0891 |
| *High Bias Limit (fem)* | -0.0624 – -0.1083 | -0.0684 – -0.1268 | -0.0664 – -0.1192 | -0.1414 – -0.2231 | -0.1015 – -0.1655 | -0.1451 – -0.2263 | -0.1438 – -0.2227 | -0.0892 – -0.1546 |
| *Extremely High Bias Limit (fem)* | Less than -0.1084 | Less than -0.1269 | Less than -0.1193 | Less than -0.2232 | Less than -0.1656 | Less than -0.2264 | Less than -0.2228 | Less than -0.1547 |
| *Low Bias Limit (mas)* | 0 – 0.0461 | 0 – 0.0502 | 0 – 0.0500 | 0 – 0.0344 | 0 – 0.0337 | 0 – 0.0347 | 0 – 0.0331 | 0 – 0.0395 |
| *Medium Bias Limit (mas)* | 0.0461 – 0.0954 | 0.0503 – 0.1016 | 0.0500 – 0.1021 | 0.0345 – 0.0783 | 0.0338 – 0.0739 | 0.0348 – 0.0795 | 0.0331 – 0.0756 | 0.0396 – 0.0866 |
| *High bias Limit (mas)* | 0.0955 – 0.1602 | 0.1017 – 0.1686 | 0.1022 – 0.1710 | 0.0784 – 0.1451 | 0.0740 – 0.1315 | 0.0796 – 0.1481 | 0.0757 – 0.1433 | 0.0867 – 0.1579 |
| *Extremely High Bias Limit (mas)* | More than 0.1603 | More than 0.1687 | More than 0.1711 | More than 0.1452 | More than 0.1316 | More than 0.1482 | More than 0.1434 | More than 0.1580 |

## Semantic Orientation Quartile Distribution for Gender-Neutral Professions in U.S. Labour Statistics

| | **Gender-Neutral Professions in Labour Statistics** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | fastText | | | GloVe | | | | word2vec |
| | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| *Judge* | -0,0474 | -0,0668 | 0,0369 | 0,0872 | 0,1172 | 0,1534 | 0,2093 | -0,0719 |
| *Artist* | 0,0135 | 0,0210 | 0,0306 | 0,1643 | 0,0687 | 0,2282 | 0,1537 | -0,0447 |
| *Photographer* | 0,0388 | 0,0350 | 0,1276 | 0,0922 | 0,0917 | 0,1379 | 0,1790 | 0,0169 |
| *Packager* | 0,0604 | -0,0140 | 0,0463 | 0,0408 | 0,1202 | -0,0816 | -0,0565 | -0,2726 |
| *Dispatcher* | 0,2042 | 0,1351 | 0,2074 | -0,0228 | 0,0899 | 0,1316 | 0,1934 | -0,0497 |
| *Statistician* | 0,0135 | 0,2044 | 0,0923 | -0,0589 | -0,0648 | 0,0242 | 0,0154 | 0,1104 |
| *Bartender* | 0,3213 | 0,1365 | 0,2132 | 0,0503 | 0,3360 | 0,1573 | 0,1915 | -0,0301 |
| *Scientist* | -0,0140 | 0,0621 | 0,0473 | 0,0712 | 0,0879 | 0,2353 | 0,2157 | -0,0157 |
| *Producer* | 0,0896 | 0,0674 | 0,0621 | 0,2088 | 0,1879 | 0,2006 | 0,1922 | -0,0892 |
| *Coach* | 0,2158 | 0,3549 | 0,4003 | 0,4580 | 0,2444 | 0,4763 | 0,5207 | 0,3775 |
| *Low Bias Limit (fem)* | 0 – -0.0287 | 0 – -0.0305 | 0 – -0.0303 | 0 – -0.0729 | 0 – -0.0502 | 0 – -0.0761 | 0 – -0.0764 | 0 – -0.0423 |
| *Medium Bias Limit (fem)* | -0.0288 – -0.0623 | -0.0306 – -0.0683 | -0.0304 – -0.0663 | -0.0730 – -0.1413 | -0.0503 – -0.1014 | -0.0762 – -0.1450 | -0.0765 – -0.1437 | -0.0424 – -0.0891 |
| *High Bias Limit (fem)* | -0.0624 – -0.1083 | -0.0684 – -0.1268 | -0.0664 – -0.1192 | -0.1414 – -0.2231 | -0.1015 – -0.1655 | -0.1451 – -0.2263 | -0.1438 – -0.2227 | -0.0892 – -0.1546 |
| *Extremely High Bias Limit (fem)* | Less than -0.1084 | Less than -0.1269 | Less than -0.1193 | Less than -0.2232 | Less than -0.1656 | Less than -0.2264 | Less than -0.2228 | Less than -0.1547 |
| *Low Bias Limit (mas)* | 0 – 0.0461 | 0 – 0.0502 | 0 – 0.0500 | 0 – 0.0344 | 0 – 0.0337 | 0 – 0.0347 | 0 – 0.0331 | 0 – 0.0395 |
| *Medium Bias Limit (mas)* | 0.0461 – 0.0954 | 0.0503 – 0.1016 | 0.0500 – 0.1021 | 0.0345 – 0.0783 | 0.0338 – 0.0739 | 0.0348 – 0.0795 | 0.0331 – 0.0756 | 0.0396 – 0.0866 |
| *High bias Limit (mas)* | 0.0955 – 0.1602 | 0.1017 – 0.1686 | 0.1022 – 0.1710 | 0.0784 – 0.1451 | 0.0740 – 0.1315 | 0.0796 – 0.1481 | 0.0757 – 0.1433 | 0.0867 – 0.1579 |
| *Extremely High Bias Limit (mas)* | More than 0.1603 | More than 0.1687 | More than 0.1711 | More than 0.1452 | More than 0.1316 | More than 0.1482 | More than 0.1434 | More than 0.1580 |

## Cosine Similarity in Adjectives for Top Extreme Masculine Profession Nouns in U.S. Labour Statistics

mason

| fastText | | | GloVe | | | | word2vec |
|---|---|---|---|---|---|---|---|
| **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| artisan (0.5459) | artisan (0.5511) | artisan (0.5576) | classic (0.3081) | artisan (0.4762) | artisan (0.3584) | artisan (0.4763) | N/A |
| masonic (0.5428) | doric (0.4400) | masonic (0.5226) | brown (0.2685) | monumental (0.3935) | bronx-born (0.2721) | masonic (0.3625) | N/A |
| worshipful (0.4384) | unglazed (0.4376) | itinerant (0.4564) | graduated (0.2648) | masonic (0.3754) | government-employed (0.2505) | skilled (0.3196) | N/A |
| monumental (0.4153) | ceramic (0.4350) | architectural (0.4494) | starred (0.2528) | skilled (0.3633) | out-of-work (0.2600) | monumental (0.2922) | N/A |
| heraldic (0.4072) | janitorial (0.4338) | brick-built (0.4287) | married (0.2437) | self-employed (0.2796) | Locomotive (0.2522) | Teutonic (0.27287) | N/A |

firefighter

| fastText | | | GloVe | | | | word2vec |
|---|---|---|---|---|---|---|---|
| **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| paramedical (0.5459) | fire-retardant (0.5374) | off-duty (0.5019) | injured (0.3799) | off-duty (0.4328) | off-duty (0.4320) | off-duty (0.4375) | irresistible (0.6204) |
| off-duty (0.5459) | wind-driven (0.5120) | fire-retardant (0.4903) | probationary (0.3088) | on-call (0.4249) | nearby (0.4034) | injured (0.4077) | miraculous (0.6089) |
| on-duty (0.5475) | fire-stricken (0.5052) | retardant (0.48842) | rescued (0.3048) | full-time (0.38196) | injured (0.4016) | dead (0.3956) | quadriplegic (0.6018) |
| on-call (0.4947) | soot-covered (0.5050) | on-duty (0.4742) | hospitalized (0.30377) | injured (0.3662) | medical (0.3336) | medical (0.3244) | non-sworn (0.5998) |
| non-sworn (0.4679) | off-duty (0.4927) | wind-driven (0.4613) | stabbed (0.2940) | on-duty (0.3645) | unable (0.3231) | local (0.3237) | N/A |

| carpenter | | | | | | | |
|---|---|---|---|---|---|---|---|
| **fastText** | | | **GloVe** | | | | **word2vec** |
| **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| artisan (0.5772) | artisan (0.5661) | artisan (0.5790) | retired (0.3336) | artisan (0.5625) | unemployed (0.4114) | artisan (0.5253) | N/A |
| semi-skilled (0.5126) | itinerant (0.5221) | itinerant (0.5459) | hired (0.3319) | skilled (0.4925) | artisan (0.4187) | unemployed (0.4373) | N/A |
| itinerant (0.5000) | self-taught (0.5140) | self-taught (0.4861) | married (0.3224) | itinerant (0.4122) | self-employed (0.3843) | skilled (0.4353) | N/A |
| skilled (0.4641) | janitorial (0.5014) | out-of-work (0.4786) | gothic (0.3019) | self-employed (0.3716) | part-time (0.3802) | itinerant (0.4237) | N/A |
| self-educated (0.4512) | weatherbeaten (0.4847) | semi-skilled (0.4670) | old (0.2940) | unemployed (0.3300) | itinerant (0.3785) | self-employed (0.4108) | N/A |

| constructor | | | | | | | |
|---|---|---|---|---|---|---|---|
| **fastText** | | | **GloVe** | | | | **word2vec** |
| **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| prototype-based (0.5113) | incident-packed (0.4717) | constructional (0.4986) | user-defined (0.3531) | user-defined (0.3340) | all-japanese (0.2776) | mathematical (0.2960) | non-null (0.7681) |
| constructional (0.5000) | series-best (0.4273) | co-driven (0.4638) | boolean (0.3262) | generic (0.2929) | all-around (0.2750) | non-major (0.2828) | tiny (0.7446) |
| Boolean (0.4441) | championship-winning (0.4264) | production-based (0.4152) | highest-ranked (0.2979) | object-oriented (0.2809) | non-major (0.2651) | user-defined (0.2814) | empty (0.7344) |
| enumerable (0.4443) | pole-winning (0.4242) | constructible (0.4131) | recursive (0.2780) | consecutive (0.2659) | series-leading (0.2417) | successive (0.2456) | boolean (0.7287) |
| co-driven (0.4286) | co-driven (0.4069) | user-defined (0.4117) | deductive (0.2720) | logical (0.2571) | consecutive (0.2365) | all-around (0.2451) | N/A |

## Cosine Similarity in Adjectives for Top Extreme Feminine Profession Nouns in U.S. Labour Statistics

| nutritionist | | | | | | | |
|---|---|---|---|---|---|---|---|
| **fastText** | | | **GloVe** | | | | **word2vec** |
| **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| nutritional (0.7559) | nutritional (0.7312) | nutritional (0.7205) | nutritional (0.3390) | nutritional (0.3679) | nutritional (0.4377) | nutritional (0.4590) | naturopathic (0.7348) |
| nutritive (0.5975) | nutritive (0.6770) | healthful (0.6476) | clinical (0.3310) | board-certified (0.3474) | dietary (0.4105) | dietary (0.4033) | macrobiotic (0.6781) |
| naturopathic (0.5821) | healthful (0.6740) | health-conscious (0.6207) | holistic (0.3127) | vegan (0.3474) | healthful (0.3801) | healthful (0.3796) | nutritional (0.6443) |
| macrobiotic (0.5777) | nutritious (0.6311) | high-fat (0.6125) | naturopathic (0.3029) | vegetarian (0.3193) | vegetarian (0.3687) | vegan (0.3714) | semi-vegetarian (0.6180) |
| nutritious (0.5694) | dietary (0.6190) | dietary (0.6099) | homeopathic (0.2999) | certified (0.3177) | vegan (0.3530) | vegetarian (0.3535) | clinic (0.6176) |

| nurse | | | | | | | |
|---|---|---|---|---|---|---|---|
| **fastText** | | | **GloVe** | | | | **word2vec** |
| **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| paramedical (0.6238) | paramedical (0.6017) | paramedical (0.6075) | medical (0.4617) | medical (0.5935) | medical (0.5656) | medical (0.5791) | paraprofessional (0.7081) |
| geriatric (0.5793) | obstetrical (0.5323) | obstetrical (0.5516) | pediatric (0.4373) | psychiatric (0.4873) | psychiatric (0.4601) | pregnant (0.5013) | geriatric (0.6985) |
| medical (0.5604) | medical (0.5176) | obstetric (0.5364) | psychiatric (0.4161) | elderly (0.4735) | pregnant (0.4545) | sick (0.4892) | obstetric (0.6477) |
| medical/surgical (0.5571) | neonatal (0.5106) | medical (0.5364) | newborn (0.3722) | pregnant (0.4528) | elderly (0.4498) | elderly (0.4649) | medical (0.6443) |
| hospital-based (0.5544) | geriatric (0.5103) | geriatric (0.5364) | clinical (0.3593) | young (0.4522) | surgical (0.4266) | psychiatric (0.4543) | neo-natal (0.6422) |

| librarian | | | | | | | |
|---|---|---|---|---|---|---|---|
| **fastText** | | | **GloVe** | | | | **word2vec** |
| **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| bibliographical (0.5551) | emerita (0.5110) | bibliographical (0.5742) | part-time (0.3252) | academic (0.4301) | retired (0.3871) | academic (0.3932) | /critic (0.6283) |
| bibliographic (0.5342) | paraprofessional (0.4988) | bibliographic (0.5344) | emerita (0.3189) | paraprofessional (0.3936) | elementary (0.3816) | part-time (0.3756) | N/A |
| paraprofessional (0.4874) | elementary (0.4975) | emerita (0.5266) | academic (0.3124) | full-time (0.3647) | academic (0.3336) | paraprofessional (0.3608) | N/A |
| secretarial (0.4780) | literary (0.4859) | Genealogical (0.5177) | honorary (0.3114) | part-time (0.3528) | part-time (0.3217) | full-time (0.3597) | N/A |
| Philological (0.4693) | salutatorian (0.4828) | Philological (0.5144) | retired (0.3064) | professional (0.3507) | literary (0.3164) | literary (0.3596) | N/A |

| veterinarian | | | | | | | |
|---|---|---|---|---|---|---|---|
| **fastText** | | | **GloVe** | | | | **word2vec** |
| **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| medical (0.5699) | animal-based (0.5413) | board-certified (0.5431) | pediatric (0.3595) | medical (0.4510) | medical (0.4226) | medical (0.4250) | pre-surgical (0.6132) |
| paramedical (0.5435) | board-certified (0.5348) | pediatric (0.5255) | naturalist (0.3419) | dental (0.4084) | infected (0.3697) | pediatric (0.3645) | geriatric (0.6108) |
| osteopathic (0.5349) | paramedical (0.5190) | naturopathic (0.5075) | forensic (0.3107) | licensed (0.3852) | clinical (0.3590) | dental (0.3587) | non-breed (0.5991) |
| pediatric (0.5272) | naturopathic (0.5191) | mixed-breed (0.5032) | medical (0.3061) | pediatric (0.3780) | pediatric (0.3595) | infected (0.3456) | naturopathic (0.5987) |
| naturopathic (0.5257) | zoological (0.5066) | paramedical (0.4959) | sick (0.3072) | board-certified (0.3607) | dental (0.3592) | sick (0.3407) | urologic (0.5978) |

## Semantic Orientation for Extreme Masculine Professions in Word Embeddings

| Extreme Masculine Professions in Word Embeddings | | | | | | | |
|---|---|---|---|---|---|---|---|
| fastText | | | GloVe | | | | word2vec |
| **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| smith (0.5915) | captain (0.5506) | quarterback (0.5766) | quarterback (0.5887) | Captain (0.4709) | groundskeeper (0.5346) | Captain (0.4998) | wizard (0.5763) |
| captain (0.5598) | cop (0.5209) | pitcher (0.5643) | player (0.5674) | shoeshine (0.4633) | wizard (0.5271) | ballplayer (0.4880) | players (0.5495) |
| manager (0.5197) | sergeant (0.5098) | actor (0.5426) | guard (0.5663) | Musketeers (0.4553) | team-owner (0.5164) | sergeant (0.4783) | N/A |
| skipper (0.5140) | thief (0.4594) | player (0.5295) | Captain (0.5477) | helmer (0.4437) | linebacker (0.4866) | linebacker (0.4705) | N/A |
| actor (0.4773) | sheriff (0.4321) | outfielder (0.5002) | officer (0.5274) | cabbie (0.4217) | quarterback (0.4284) | Colonel (0.4623) | N/A |
| player (0.4651) | Colonel (0.4317) | captain (0.4949) | coach (0.5207) | manservant (0.4196) | cornerback (0.4090) | leftfielder (0.4529) | N/A |
| coach (0.4580) | actor (0.4293) | guard (0.4884) | manager (0.5160) | cowboy (0.4126) | Manager (0.4076) | buffoon (0.4404) | N/A |
| quarterback (0.4361) | detective (0.4209) | driver (0.4871) | Colonel (0.5137) | Mafioso (0.4090) | coach-general (0.3965) | General-Major (0.4381) | N/A |
| linebacker (0.4334) | drummer (0.4166) | skipper (0.4738) | builder (0.5073) | Sargeant (0.4067) | player (0.3943) | skipper (0.4201) | N/A |
| drummer (0.4243) | policeman (0.4090) | officer (0.4595) | actor (0.4985) | Gladiador (0.4029) | N/A | centreback (0.4126) | N/A |

## Semantic Orientation for Extreme Feminine Professions in Word Embeddings

| Extreme Feminine Professions in Word Embeddings | | | | | | | |
|---|---|---|---|---|---|---|---|
| fastText | | | GloVe | | | | word2vec |
| **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| chanteuse (-0.6311) | abbess (-0.4747) | model (-0.6722) | actress (-0.6563) | deaconess (-0.4056) | chairperson (-0.5282) | nurse-midwife (-0.4816) | chanteuse (-0.7899) |
| abbess (-0.6141) | model (-0.4502) | chairperson (-0.6672) | vedette (-0.6274) | alumna (-0.4006) | actress (-0.4793) | alumna (-0.4713) | songstress (-0.7164) |
| alumna (-0.6107) | actress (-0.4383) | supermodel (-0.6498) | chairperson (-0.6166) | mayoress (-0.3833) | dominatrix (-0.4481) | actress (-0.4365) | housewife (-0.7153) |
| vedette (-0.6011) | songstress (-0.4216) | poetess (-0.6159) | poetess (-0.5731) | abbess (-0.3781) | nurse-midwife (-0.4390) | model (-0.4360) | actress (-0.7151) |
| comedienne (-0.5925) | alumna (-0.4207) | actress (-0.5663) | alumna (-0.5647) | diva (-0.3377) | supermodel (-0.4368) | singer (-0.4334) | alumna (-0.7016) |
| sculptress (-0.5889) | huntress (-0.4200) | ex-model (-0.5228) | abbess (-0.5319) | benefactress (-0.3293) | diva (-0.4349) | benefactress (-0.4295) | comedienne (-0.7015) |
| ballerina (-0.5850) | singer (-0.4003) | general-manager (-0.5170) | ballerina (-0.5279) | nun (-0.3242) | chanteuse (-0.4335) | midwife (-0.4290) | showgirl (-0.6665) |
| songstress (-0.5808) | headmistress (-0.3956) | N/A | chanteuse (-0.5233) | nurse (-0.3204) | model (-0.4275) | hostess (-0.4066) | hostess (-0.6619) |
| mezzo-soprano (-0.5798) | N/A | N/A | patroness (-0.5141) | midwife (-0.3087) | shepperdess (-0.4218) | chanteuse (-0.3860) | Homemaker (-0.6612) |
| contralto (-0.5509) | N/A | N/A | prioress (-0.5070) | handmaid (-0.3055) | singer (-0.4145) | songstress (-0.3842) | nurse_midwife (-0.6529) |

**Cosine Similarity in Adjectives for Top Extreme Masculine Profession Nouns in Word Embedding and Their Female Counterparts (When Available)**

| captain | | | | | | | |
|---|---|---|---|---|---|---|---|
| **fastText** | | | **GloVe** | | | | **word2vec** |
| **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| second-in-command (0.5474) | cup-winning (0.609) | cup-winning (0.5809) | named (0.3949) | senior (0.5250) | injured (0.4169) | stand-in (0.4164) | –corporal (0.7293) |
| commanding (0.5144) | non-playing (0.5924) | most-capped (0.5797) | returned (0.3809) | commanding (0.4795) | one-day (0.3873) | veteran (0.4007) | commanding (0.7164) |
| second-choice (0.5112) | cup-winner (0.5831) | out-of-form (0.5581) | replaced (0.3618) | youngest (0.4500) | experienced (0.3848) | injured (0.3974) | pseudo-nihilistic (0.6657) |
| top-score (0.5055) | out-of-form (0.5775) | ashes-winning (0.5491) | commanding (0.3607) | victorious (0.4227) | in-form (0.3848) | experienced (0.3921) | nicosian (0.6535) |
| all-rounder (0.4961) | most-capped (0.5759) | cup-winner (0.5359) | retired (0.3451) | experienced (0.4221) | latter (0.4031) | all-rounder (0.3716) | N/A |

| player | | | | | | | |
|---|---|---|---|---|---|---|---|
| **fastText** | | | **GloVe** | | | | **word2vec** |
| **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| professional (0.4763) | all-star (0.5550) | all-star (0.5424) | professional (0.5322) | select (0.5916) | professional (0.5537) | professional (0.5486) | ultra-athletic (0.6821) |
| overage (0.4693) | collegian (0.5492) | preseason (0.506) | best (0.5000) | professional (0.5715) | best (0.5357) | select (0.5089) | co-special (0.6468) |
| all-star (0.4691) | preseason (0.5402) | postseason (0.5057) | defensive (0.4650) | defensive (0.5048) | talented (0.5267) | great (0.4835) | most-valuable (0.6456) |
| championship-winning (0.4661) | postseason (0.5295) | professional (0.4789) | veteran (0.4564) | able (0.5023) | great (0.5168) | talented (0.4825) | positionless (0.6353) |
| in-game (0.4569) | all-rookie (0.5199) | coachable (0.4785) | talented (0.4446) | notable (0.4677) | better (0.5026) | able (0.4791) | undroppable (0.6348) |

actor

| fastText | | | GloVe | | | | word2vec |
|---|---|---|---|---|---|---|---|
| **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| comedic (0.5751) | Oscar-winner (0.6732) | oscar-winner (0.6366) | nominated (0.4573) | best (0.5512) | comic (0.5176) | best (0.5114) | –nominated (0.7636) |
| directorial (0.5261) | oscar-nominated (0.6506) | oscar-winning (0.6249) | comedic (0.4428) | well-known (0.5095) | oscar-winning (0.5173) | oscar-winning (0.4963) | directorial (0.7300) |
| oscar-nominated (0.5155) | emmy-winning (0.6279) | oscar-nominated (0.6155) | supporting (0.4247) | notable (0.5030) | best (0.4978) | young (0.4875) | oscar-nominated (0.7136) |
| tony-nominated (0.5148) | directorial (0.6183 | comedic (0.6119) | acclaimed (0.4201) | veteran (0.5017) | young (0.4952) | award-winning (0.4795) | comedic (0.7081) |
| shakespearean (0.5116) | emmy-nominated (0.6155) | directorial (0.6031) | award-winning (0.4100) | theatrical (0.4949) | onstage (0.4642) | opposite (0.4720) | N/A |

actress

| fastText | | | GloVe | | | | word2vec |
|---|---|---|---|---|---|---|---|
| **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| oscar-winning (0.4890) | oscar-winner (0.6888) | oscar-winner (0.6232) | nominated (0.4444) | best (0.5361) | oscar-winning (0.5234) | oscar-winning (0.4890) | –nominated (0.7111) |
| award-winning (0.4887) | oscar-nominated (0.6619) | oscar-nominated (0.6102) | best (0.4292) | aspiring (0.4950) | glamorous (0.4688) | award-winning (0.4887) | best- (0.6865) |
| comedian (0.4868) | oscar-winning (0.6376) | award-nominated (0.6046) | née (0.4101) | née (0.4461) | award-winning (0.4635) | aspiring (0.4812) | co-nominated (0.6637) |
| aspiring (0.4812) | award-nominated (0.6123) | oscar-winning (0.5876) | blonde (0.3944) | young (0.4453) | oscar-nominated (0.4621) | oscar-nominated (0.4433) | N/A |
| best (0.4804) | emmy-nominated (0.6033) | globe-winning (0.5684) | glamorous (0.3870) | outstanding (0.4417) | best (0.4525) | young (0.4623) | N/A |

| quarterback | | | | | | | |
|---|---|---|---|---|---|---|---|
| **fastText** | | | **GloVe** | | | | **word2vec** |
| **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| defensive (0.5488) | bowl-winning (0.6887) | pass-rushing (0.6974) | defensive (0.5573) | defensive (0.5576) | defensive (0.5478 | defensive (0.5488) | defensive (0.8463) |
| second-year (0.5203) | pass-oriented (0.6536) | pass-catching (0.696) | offensive (0.4411) | offensive (0.4781) | preseason (0.4885 | preseason (0.501) | co-offensive (0.8184) |
| preseason (0.5010) | run-oriented (0.6504) | )pass-oriented (0.6479) | preseason (0.4232) | preseason (0.4324) | offensive (0.4436 | offensive (0.4562) | preseason (0.7818 |
| offensive (0.4562) | nfl-best (0.6453) | nfl-best (0.6438) | undrafted (0.3648) | undefeated (0.3977) | postseason (0.4409 | postseason (0.4249) | pro-style (0.7724) |
| postseason (0.4249) | undrafted (0.6421) | run-oriented (0.6375) | postseason (0.3375) | junior (0.3855) | offseason (0.4238 | offseason (0.4008) | pro-bowl (0.7700) |

**Cosine Similarity in Adjectives for Top Extreme Feminine Profession Nouns in Word Embedding and Their Male Counterparts (When Available)**

alumna

| fastText | | | GloVe | | | | word2vec |
|---|---|---|---|---|---|---|---|
| **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| salutatorian (0.5139) | pre-med (0.5142) | salutatorian (0.5224) | emerita (0.3956) | distinguished (0.4209) | distinguished (0.2871) | distinguished (0.4306) | emerita (0.7349) |
| emerita (0.5064) | high-achieving (0.4948) | emerita (0.5119) | distinguished (0.3932) | illustrious (0.3288) | varsity (0.285) | illustrious (0.3393) | salutatorian (0.7134) |
| high-achieving (0.468) | co-ed (0.4899) | pre-med (0.5003) | eighteen-year-old (0.3408) | esteemed (0.2901) | esteemed (0.2835) | esteemed (0.3077) | pre-doctoral (0.6693) |
| college-bound (0.4672 | salutatorian (0.4804) | pre-law (0.4985) | estimable (0.3166) | asian-american (0.2862) | illustrious (0.2794) | emerita (0.2917) | montessorian (0.6445) |
| university-based (0.4663 | co-educational (0.4766) | co-ed (0.4950) | vivacious (0.3163) | honorary (0.2859) | high-achieving (0.2788) | well-liked (0.2859) | N/A |

alumnus

| fastText | | | GloVe | | | | word2vec |
|---|---|---|---|---|---|---|---|
| **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| extracurricular (0.4924) | varsity (0.5343) | extracurricular (0.5233) | distinguished (0.4621) | distinguished (0.5686) | distinguished (0.4462) | distinguished (0.5457) | emeritus (0.7533) |
| philanthropic (0.4821) | high-achieving (0.5293) | co-curricular (0.5202) | emeritus (0.4551) | notable (0.5155) | athletic (0.4260) | notable (0.4927) | honorary (0.7161) |
| distinguished (0.4633) | first-year (0.5073) | distinguished (0.5127) | honorary (0.4453) | long-time (0.4658) | academic (0.4240) | prestigious (0.4517) | emerita (0.6568) |
| high-achieving (0.4586) | extracurricular (0.5012) | first-year (0.5123) | esteemed (0.4078) | prestigious (0.4477) | prestigious (0.3657) | athletic (0.4389) | distinguished (0.6544) |
| varsity (0.4505) | distinguished (0.4816) | salutatorian (0.4907) | honored (0.3883) | prominent (0.4264) | longtime (0.3653) | first-year (0.3965) | ex-athletic (0.6483) |

| actress | | | | | | | |
|---|---|---|---|---|---|---|---|
| **fastText** | | | **GloVe** | | | | **word2vec** |
| **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| oscar-winning (0.4890) | oscar-winner (0.6888) | oscar-winner (0.6232) | nominated (0.4444) | best (0.5361) | oscar-winning (0.5234) | oscar-winning (0.4890) | –nominated (0.7111) |
| award-winning (0.4887) | oscar-nominated (0.6619) | oscar-nominated (0.6102) | best (0.4292) | aspiring (0.4950) | glamorous (0.4688) | award-winning (0.4887) | best- (0.6865) |
| comedian (0.4868) | oscar-winning (0.6376) | award-nominated (0.6046) | née (0.4101) | née (0.4461) | award-winning (0.4635) | aspiring (0.4812) | co-nominated (0.6637) |
| aspiring (0.4812) | award-nominated (0.6123) | oscar-winning (0.5876) | blonde (0.3944) | young (0.4453) | oscar-nominated (0.4621) | oscar-nominated (0.4433) | N/A |
| best (0.4804) | emmy-nominated (0.6033) | globe-winning (0.5684) | glamorous (0.3870) | outstanding (0.4417) | best (0.4525) | young (0.4623) | N/A |

| actor | | | | | | | |
|---|---|---|---|---|---|---|---|
| **fastText** | | | **GloVe** | | | | **word2vec** |
| **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| comedic (0.5751) | Oscar-winner (0.6732) | oscar-winner (0.6366) | nominated (0.4573) | best (0.5512) | comic (0.5176) | best (0.5114) | –nominated (0.7636) |
| directorial (0.5261) | oscar-nominated (0.6506) | oscar-winning (0.6249) | comedic (0.4428) | well-known (0.5095) | oscar-winning (0.5173) | oscar-winning (0.4963) | directorial (0.7300) |
| oscar-nominated (0.5155) | emmy-winning (0.6279) | oscar-nominated (0.6155) | supporting (0.4247) | notable (0.5030) | best (0.4978) | young (0.4875) | oscar-nominated (0.7136) |
| tony-nominated (0.5148) | directorial (0.6183 | comedic (0.6119) | acclaimed (0.4201) | veteran (0.5017) | young (0.4952) | award-winning (0.4795) | comedic (0.7081) |
| shakespearean (0.5116) | emmy-nominated (0.6155) | directorial (0.6031) | award-winning (0.4100) | theatrical (0.4949) | onstage (0.4642) | opposite (0.4720) | N/A |

| chanteuse | | | | | | | |
|---|---|---|---|---|---|---|---|
| **fastText** | | | **GloVe** | | | | **word2vec** |
| **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| fatale (0.6258) | chart-topper (0.6462) | chart-topping (0.6299) | boozy (0.4018) | elusive (0.558) | world-weary (0.3463) | wagnerian (0.3671) | fatale (0.7712) |
| glamorous (0.5681) | grammy-winner (0.6434) | sassy (0.6273) | soulful (0.3804) | sultry (0.3392) | soulful (0.3307) | sultry (0.3627) | désenchantée (0.7603) |
| operatic (0.5303) | chart-topping (0.6355) | fatale (0.6241) | inflected (0.3709) | seule (0.3132) | melancholic (0.3267) | soulful (0.3596) | souriante (0.7372) |
| sassy (0.5521) | grammy-winning (0.6260) | sultry (0.6176) | saucy (0.3709) | puerto-rican (0.3000) | parisian (0.3209) | mellifluous (0.3360) | chérie (0.7293) |
| seductive (0.5495) | sassy (0.6110) | seductive (0.6083) | sultry (0.3705) | coquettish (0.2889) | mellifluous (0.3081) | parisian (0.3178) | démodé (0.7093) |

| chanteur | | | | | | | |
|---|---|---|---|---|---|---|---|
| **fastText** | | | **GloVe** | | | | **word2vec** |
| **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| N/A | N/A | N/A | N/A | N/A | N/A | N/A | noyeé (0.8627) |
| N/A | N/A | N/A | N/A | N/A | N/A | N/A | souriante (0.8627) |
| N/A | N/A | N/A | N/A | N/A | N/A | N/A | l'aveugle (0.8610) |
| N/A | N/A | N/A | N/A | N/A | N/A | N/A | infidèle (0.8608) |
| N/A | N/A | N/A | N/A | N/A | N/A | N/A | jaloux (0.8588) |

| abbess | | | | | | | |
|---|---|---|---|---|---|---|---|
| **fastText** | | | **GloVe** | | | | **word2vec** |
| **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| cistercian (0.6348) | benedictine (0.5753) | benedictine (0.7349) | benedictine (0.5481) | benedictine (0.5539) | benedictine (0.3601) | benedictine (0.5806) | carmelite (0.7470) |
| carmelite (0.6143) | monastic (0.5448) | cistercian (0.7195) | cistercian (0.5220) | cistercian (0.5388) | flaxen-haired (0.3285) | cistercian (0.5440) | cistercian (0.7329) |
| monastic (0.6071) | cistercian (0.5394) | monastic (0.6766) | carmelite (0.4833) | carmelite (0.4595) | stanford-educated (0.3232) | carmelite (0.5358) | premonstratensian (0.7213) |
| premonstratensian (0.6020) | carmelite (0.5325) | augustinian (0.6728) | augustinian (0.4280) | augustinian (0.4434) | less-powerful (0.3082) | augustinian (0.4666) | benedictine (0.7093) |
| conventual (0.6000) | taoist (0.5217) | carmelite (0.6711) | monastic (0.3948) | monastic (0.4053) | augustinian (0.3055) | monastic (0.4143) | N/A |

| abbot | | | | | | | |
|---|---|---|---|---|---|---|---|
| **fastText** | | | **GloVe** | | | | **word2vec** |
| **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| monastic (0.7361) | monastic (0.6135) | benedictine (0.7701) | benedictine (0.6206) | Benedictine (0.6573) | Buddhist (0.4166) | cistercian (0.6385) | tironensian (0.8060) |
| benedictine (0.7346) | taoist (0.5839) | monastic (0.7635) | cistercian (0.5631) | cistercian (0.6489) | franciscan (0.4150) | Benedictine (0.6280) | benedictine (0.7699) |
| cistercian (0.6861) | buddhist (0.5738) | cistercian (0.7342) | monastic (0.4950) | monastic (0.6004) | Shaolin (0.3168) | monastic (0.5848) | cistercian (0.7691) |
| premonstratensian (0.6643) | benedictine (0.5656) | premonstratensian (0.7202) | augustinian (0.4727) | augustinian (0.5040) | Benedictine (0.3768) | augustinian (0.5425) | premonstratensian (0.7679) |
| ecclesiastic (0.6490) | shaolin (0.5649) | augustinian (0.6999) | franciscan (0.4596) | ecclesiastic (0.4461) | taoist (0.3567) | franciscan (0.4753) | celestines (0.7627) |

| model | | | | | | | |
|---|---|---|---|---|---|---|---|
| **fastText** | | | **GloVe** | | | | **word2vec** |
| **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| top-of-the-range (0.5558) | performance-oriented (0.5819) | sportier (0.5598) | same (0.4099) | similar (0.5482) | new (0.5216) | similar (0.5311) | geoadditive (0.7629) |
| top-of-the-line (0.5372) | full-sized (0.5737) | high-performance (0.5460) | hybrid (0.4084) | different (0.5353) | standard (0.5108) | different (0.5157) | thurstonian (0.725) |
| limited-production (0.5238) | sporty (0.5687) | top-of-the-line (0.5419) | mathematical (0.3782) | available (0.4965) | similar (0.4918) | standard (0.5021) | microfounded (0.6934) |
| autoregressive (0.5210) | sleekest (0.5584) | full-sized (0.5400) | larger (0.3773) | new (0.4938) | different (0.4794) | larger (0.4834) | multinominal (0.6821) |
| semi-empirical (0.5177) | top-of-the-line (0.5476) | eight-speed (0.5394) | dynamic (0.3768) | fit (0.4887) | hybrid (0.4732) | new (0.4691) | all-electric (0.6811) |

| songstress | | | | | | | |
|---|---|---|---|---|---|---|---|
| **fastText** | | | **GloVe** | | | | **word2vec** |
| **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| sultry (0.6148 | soulful (0.6297 | grammy-nominated (0.6521 | sultry (0.4445 | girlish (0.3515 | sultry (0.4097 | sultry (0.4506 | soulful (0.7526 |
| soulful (0.6007 | grammy-nominated (0.6237 | soulful (0.6456 | blue-eyed (0.4023 | nubile (0.347 | soulful (0.342 | grammy-winning (0.3913 | vampish (0.7017 |
| sassy (0.5767 | self-titled (0.6125 | grammy-winning (0.6347 | soulful (0.3781 | independent-minded (0.3369 | wagnerian (0.3222 | soulful (0.3558 | kissless (0.6834 |
| seductive (0.5445 | multiplatinum (0.6045 | chart-topper (0.612 | red-haired (0.3656 | self-assured (0.3305 | chirpy (0.3095 | grammy-nominated (0.3434 | neosupervital (0.6824 |
| catchy (0.5404 | breathy (0.5958 | sultry (0.6067 | seductive (0.365 | angsty (0.33 | raven-haired (0.2985 | chart-topping (0.3388 | anthemic (0.6791 |

## Semantic Orientation for Extreme Masculine Words Ending in -Man in Word Embeddings

| Extreme Masculine Words Ending in -Man in Word Embeddings | | | | | | | |
|---|---|---|---|---|---|---|---|
| fastText | | | GloVe | | | | word2vec |
| **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| Englishman (0.5841) | henchman (0.5212) | spokesman (0.5456) | spokesman (0.5628) | con-man (0.5935) | journeyman (0.5970) | journeyman (0.5965) | journeyman (0.6615) |
| batsman (0.4696) | gentleman (0.4521) | countryman (0.4799) | salesman (0.4906) | henchman (0.5721) | wingman (0.5804) | countryman (0.5898) | countryman (0.6288) |
| salesman (0.4486) | astro-man (0.4382) | Englishman (0.4734) | journeyman (0.4906) | hitman (0.5356) | countryman (0.5709) | watchman (0.5742) | Patrolman (0.5924) |
| Frenchman (0.4480) | snowman (0.4303) | showman (0.4515) | Englishman (0.4749) | repairman (0.5167) | big-man (0.5299) | Englishman (0.5557) | iceman (0.5496) |
| Spiderman (0.4293) | policeman (0.4090) | lineman (0.4408) | countryman (0.4490) | foreman (0.5101) | Englishman (0.5129) | patrolman (0.5006) | lineman (0.5433) |
| journeyman (0.4107) | gunman (0.4032) | Dutchman (0.4385) | gentleman (0.4381) | headsman (0.5042) | Dutchman (0.5048) | Dutchman (0.4975) | N/A |
| defenseman (0.4057) | swingman (0.3905) | journeyman (0.4229) | lineman (0.4371) | woodsman (0.4824) | groundsman (0.4817) | henchman (0.4941) | N/A |
| chairman (0.3862) | Batman (0.3892) | batsman (0.4204) | foreman (0.3996) | countryman (0.4709) | madman (0.4728) | utilityman (0.4902) | N/A |
| foreman (0.3844) | Spider-Man (0.3790) | salesman (0.4133) | N/A | Frontman (0.4563) | fieldsman (0.4634) | linesman (0.4881) | N/A |
| spokesman (0.3837) | foreman (0.3689) | defenseman (0.4128) | N/A | Iceman (0.4332) | Birdman (0.4586) | woodsman (0.4870) | N/A |

## Semantic Orientation for Extreme Feminine Words Ending in -Woman in Word Embeddings

| Extreme Feminine Professions in Word Embeddings | | | | | | | |
|---|---|---|---|---|---|---|---|
| fastText | | | GloVe | | | | word2vec |
| **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| businesswoman (0.7352) | businesswoman (0.5779) | chairwoman (0.7341) | Chairwoman (0.8093) | businesswoman (0.4822) | chairwoman (0.7480) | Chairwoman (0.6461) | frontwoman (0.8248) |
| assemblywoman (0.7013) | chairwoman (0.5337) | Congresswoman (0.5898) | Congresswoman (0.7229) | laywoman (0.4731) | businesswoman (0.5651) | businesswoman (0.5731) | chairwoman (0.8112) |
| chairwoman (0.6948) | Assemblywoman (0.4745) | Spokeswoman (0.5104) | businesswoman (0.6040) | chairwoman (0.4601) | Frenchwoman (0.5566) | congresswoman (0.4806) | businesswoman (0.7654) |
| congresswoman (0.6898) | noblewoman (0.4681) | N/A | co-chairwoman (0.5379) | congresswoman (0.4401) | congresswoman (0.5500) | vice-chairwoman (0.4611) | countrywoman (0.6784) |
| countrywoman (0.6584) | Congresswoman (0.4680) | N/A | Spokeswoman (0.5372) | Assemblywoman (0.3961) | councilwoman (0.5396) | frontwoman (0.4473) | Congresswoman (0.6553) |
| councilwoman (0.6442) | frontwoman (0.3961) | N/A | frontwoman (0.5296) | servicewoman (0.3624) | assemblywoman (0.5391) | Frenchwoman (0.4462) | councilwoman (0.6176) |
| committeewoman (0.6029) | N/A | N/A | Assemblywoman (0.5282) | spokeswoman (0.3497) | anchorwoman (0.5316) | countrywoman (0.4448) | spokeswoman (0.6122) |
| gentlewoman (0.5830) | N/A | N/A | countrywoman (0.5206) | councilwoman (0.3416) | spokeswoman (0.5303) | laywoman (0.4384) | N/A |
| noblewoman (0.5485) | N/A | N/A | Councilwoman (0.5197) | frontwoman (0.3165) | countrywoman (0.5059) | councilwoman (0.4290) | N/A |
| newswoman (0.5269) | N/A | N/A | noblewoman (0.5166) | committeewoman (0.2999) | committeewoman (0.4361) | spokeswoman (0.4162) | N/A |

**Cosine Similarity in Adjectives for 'Man'-Ending Profession Nouns in Word Embedding and their 'Woman'- and 'Person'-Ending Counterparts**

| countryman | | | | | | | |
|---|---|---|---|---|---|---|---|
| **fastText** | | | **GloVe** | | | | **word2vec** |
| **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| magnanimous (0.4721) | unheralded (0.4189) | newly-crowned (0.4701) | unseeded (0.5661) | beloved (0.3502) | unseeded (0.4188) | unseeded (0.3956) | ex–grand (0.6402) |
| despaired (0.4615) | highly-fancied (0.4053) | unheralded (0.4340) | seeded (0.5027) | experienced (0.3412) | upset (0.3988) | upset (0.3850) | superb (0.634) |
| chivalrous (0.4292) | recently-crowned (0.4035) | fancied (0.4244) | unheralded (0.4194) | like-minded (0.3378) | favourite (0.3432) | unheralded (0.3618) | N/A |
| valiant (0.4272) | dispossessed (0.4008) | dispossessed (0.4243) | aussie (0.3140) | veteran (0.3365) | unheralded (0.3308) | favourite (0.3401) | N/A |
| disheartened (0.4264) | powerfully-built (0.3989) | unfancied (0.4145) | sordo (0.2925) | grateful (0.3319) | ordinary (0.3229) | argentinian (0.3368) | N/A |

| countrywoman | | | | | | | |
|---|---|---|---|---|---|---|---|
| **fastText** | | | **GloVe** | | | | **word2vec** |
| **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| unseeded (0.5550) | unseeded (0.5883) | unseeded (0.5982) | unseeded (0.5210) | unseeded (0.3976) | unseeded (0.4672) | unseeded (0.4950) | N/A |
| top-seeded (0.5148) | personal-best (0.5408) | newly-crowned (0.5413) | second-placed (0.2961) | non-adjacent (0.3215) | top-seeded (0.4038) | top-seeded (0.4217) | N/A |
| lowest-seeded (0.4858) | newly-crowned (0.5143) | world-ranked (0.5307) | N/A | Juridical (0.3067) | top-ranked (0.3094) | top-ranked (0.3294) | N/A |
| pre-olympic (0.4815) | world-ranked (0.5116) | top-seeded (0.5237) | N/A | top-seeded (0.2989) | postdoctoral (0.2974) | in-form (0.3193) | N/A |
| in-form (0.4780) | top-seeded (0.5112) | personal-best (0.5152) | N/A | lower-ranked (0.2920) | now-retired (0.2826) | on-form (0.3176) | N/A |

| countryperson | | | | | | | |
|---|---|---|---|---|---|---|---|
| **fastText** | | | **GloVe** | | | | **word2vec** |
| **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |

journeyman

| fastText | | | GloVe | | | | word2vec |
|---|---|---|---|---|---|---|---|
| **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| self-educated (0.4836) | oft-injured (0.6210) | oft-injured (0.5558) | veteran (0.4430) | skilled (0.3900) | left-hander (0.4507) | right-hander (0.4343) | multi-skilled (0.6714) |
| skilled (0.4824) | injury-prone (0.5830) | injury-prone (0.5328) | profession al (0.3313) | itinerant (0.3837) | right-hander (0.4367) | left-hander (0.4023) | semi-retired (0.6573) |
| injury-prone (0.4745) | undrafted (0.5795) | injury-plagued (0.5316) | undrafted (0.3209) | experienced (0.3519) | veteran (0.4037) | self-taught (0.3750) | cost-controlled (0.6393) |
| well-paid (0.4702) | injury-plagued (0.5755) | left-hander (0.5248) | self-taught (0.2995) | artisan (0.3475) | pro (0.3930) | pro (0.3676) | geek (0.5853) |
| self-employed (0.4606) | left-handed (0.5688) | talented (0.5146) | semi-retired (0.2897) | unskilled (0.3445) | left-handed (0.3815) | undrafted (0.3644) | ex-offensive (0.5787) |

journeywoman

| fastText | | | GloVe | | | | word2vec |
|---|---|---|---|---|---|---|---|
| **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| N/A | N/A | N/A | N/A | N/A | N/A | N/A | workholic (0.7810) |
| N/A | N/A | N/A | N/A | N/A | N/A | N/A | sex-deprived (0.7403) |
| N/A | N/A | N/A | N/A | N/A | N/A | N/A | post-competitive (0.7397) |
| N/A | N/A | N/A | N/A | N/A | N/A | N/A | hungry (0.7267) |
| N/A | N/A | N/A | N/A | N/A | N/A | N/A | quasi-incestuous (0.7206) |

journeyperson

| fastText | | | GloVe | | | | word2vec |
|---|---|---|---|---|---|---|---|
| **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| N/A | N/A | N/A | N/A | N/A | N/A | N/A | para-professional (0.7381) |
| N/A | N/A | N/A | N/A | N/A | N/A | N/A | civil/environmental (0.7233) |
| N/A | N/A | N/A | N/A | N/A | N/A | N/A | –technical-occupational (0.7132) |
| N/A | N/A | N/A | N/A | N/A | N/A | N/A | /vocational (0.7026) |
| N/A | N/A | N/A | N/A | N/A | N/A | N/A | semi-senior (0.6996) |

Englishman

| fastText | | | GloVe | | | | word2vec |
|---|---|---|---|---|---|---|---|
| Wikipedia | Gigawords | Wikipedia + Gigawords | Original | Wikipedia | Gigawords | Wikipedia + Gigawords | Original |
| oxford-educated (0.5084) | english-born (0.5393) | all-british (0.5064) | affable (0.3602) | wealthy (0.4149) | in-form (0.3415) | in-form (0.3724) | ex-alcoholic (0.7275) |
| well-born (0.4629) | all-british (0.5301) | london-born (0.4882) | unassuming (0.3039) | young (0.4041) | british (0.3454) | english (0.3516) | philadelphian (0.7239) |
| london-born (0.4478) | glaswegian (0.5032) | liverpudlian (0.4856) | old (0.3033) | fellow (0.4038) | soft-spoken (0.3359) | british (0.3494) | aberdonian (0.7169) |
| non-british (0.4404) | liverpool-born (0.5007) | claret (0.4796) | handsome (0.2901) | middle-aged (0.3476) | favourite (0.3349) | fellow (0.3480) | impecunious (0.7144) |
| wealthy (0.4664) | england-born (0.4930) | mancunian (0.4791) | unbeaten (0.2892) | handsome (0.3462) | fellow (0.3289) | favourite (0.3269) | redoubtable (0.6997) |

Englishwoman

| fastText | | | GloVe | | | | word2vec |
|---|---|---|---|---|---|---|---|
| Wikipedia | Gigawords | Wikipedia + Gigawords | Original | Wikipedia | Gigawords | Wikipedia + Gigawords | Original |
| oxford-educated (0.5370) | well-bred (0.5061) | well-bred (0.5405) | manhattanite (0.4147) | free-spirited (0.3528) | well-bred (0.3460) | well-bred (0.4150) | impecunious (0.7387) |
| shrewish (0.5067) | prim (0.5056) | shrewish (0.4965) | strong-willed (0.3275) | vivacious (0.3479) | upper-class (0.3332) | upper-class (0.3730) | semi-invalid (0.7119) |
| amorous (0.5020) | well-born (0.5048) | vivacious (0.4894) | suburbanite (0.3196) | sweet-natured (0.3020) | prim (0.3279) | prim (0.3110) | shrewish (0.6881) |
| married (0.4892) | epistolary (0.4965) | proto-feminist (0.4878) | upper-class (0.3137) | danish-born (0.2991) | beauteous (0.3178) | headstrong (0.3073) | sexagenarian (0.6871) |
| vivacious (0.4862) | aristocratic (0.4910) | prim (0.4852) | vivacious (0.3162) | well-educated (0.2959) | raven-haired (0.2722) | beauteous (0.3063) | octogenarian (0.6862) |

Englishperson

| fastText | | | GloVe | | | | word2vec |
|---|---|---|---|---|---|---|---|
| Wikipedia | Gigawords | Wikipedia + Gigawords | Original | Wikipedia | Gigawords | Wikipedia + Gigawords | Original |
| N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |

| foreman | | | | | | | |
|---|---|---|---|---|---|---|---|
| **fastText** | | | **GloVe** | | | | **word2vec** |
| **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| janitorial (0.5055) | barrel-chested (0.4135) | janitorial (0.4291) | unanimous (0.3157) | artisan (0.3028) | retired (0.2852) | vacant (0.2882) | graziose (0.6697) |
| semi-skilled (0.4313) | janitorial (0.4061) | barrel-chested (0.4096) | vacant (0.2399) | skilled (0.3013) | vacant (0.2491) | unanimous (0.2882) | N/A |
| non-unionized (0.4130) | hard-hatted (0.3840) | out-of-work (0.3783) | undisputed (0.2345) | metallurgical (0.2965) | unanimous (0.2437) | retired (0.2640) | N/A |
| menial (0.4076) | weatherbeaten (0.3734) | semi-retired (0.3714) | greasy (0.1890) | semi-skilled (0.2609) | janitorial (0.2374) | menial (0.2439) | N/A |
| better-paying (0.3881) | pugilistic (0.3694) | menial (0.3679) | menial (0.1881) | convinced (0.2384) | monotonous (0.2247) | artisan (0.2343) | N/A |

| forewoman | | | | | | | |
|---|---|---|---|---|---|---|---|
| **fastText** | | | **GloVe** | | | | **word2vec** |
| **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| N/A | second-degree (0.5146) | not-guilty (0.5577) | impassive (0.2707) | N/A | not-guilty (0.2661) | not-guilty (0.2847) | cross-examined (0.7043) |
| N/A | first-degree (0.4995) | first-degree (0.5308) | semiconscious (0.2637) | N/A | maltese-registered (0.2645) | kingdom-based (0.2489) | N/A |
| N/A | not-guilty (0.4636) | second-degree (0.5250) | youngest-ever (0.2524) | N/A | much-quoted (0.2548) | pre-assembled (0.2395) | N/A |
| N/A | Superior (0.4558) | guilty (0.4710) | drinkable (0.2474) | N/A | greek-flagged (0.2490) | comprehensible (0.2356) | N/A |
| N/A | first-grade (0.4288) | evidentiary (0.4337) | undrinkable (0.2474) | N/A | all-white (0.2345) | much-beloved (0.2326) | N/A |

| foreperson | | | | | | | |
|---|---|---|---|---|---|---|---|
| **fastText** | | | **GloVe** | | | | **word2vec** |
| **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| N/A | not-guilty (0.5198) | N/A | N/A | N/A | all-rookie (0.2858) | N/A | grievant (0.7170) |
| N/A | evidentiary (0.4937) | N/A | N/A | N/A | all-white (0.2805) | N/A | non-testifying (0.6625) |
| N/A | second-degree (0.4875) | N/A | N/A | N/A | trans-fat-free (0.2704) | N/A | non-judicial (0.6561) |
| N/A | first-degree (0.4832) | N/A | N/A | N/A | maltese-registered (0.2685) | N/A | judicial (0.6429) |
| N/A | unsworn (0.4776) | N/A | N/A | N/A | ecclesiastic (0.2672) | N/A | pre-disciplinary (0.6356) |

lineman

| fastText | | | GloVe | | | | word2vec |
|---|---|---|---|---|---|---|---|
| Wikipedia | Gigawords | Wikipedia + Gigawords | Original | Wikipedia | Gigawords | Wikipedia + Gigawords | Original |
| defensive (0.6518) | defensive (0.7272) | defensive (0.6840) | defensive (0.6665) | defensive (0.6262) | defensive (0.6887) | defensive (0.6639) | defensive (0.8479) |
| co-defensive (0.6507) | undrafted (0.6706) | co-defensive (0.6422) | offensive (0.5444) | offensive (0.5343) | All-Pro (0.5583) | All-Pro (0.5305) | co-offensive (0.8294) |
| co-offensive (0.5822) | co-defensive (0.6417) | undrafted (0.6018) | undrafted (0.4326) | Valuable (0.4341) | offensive (0.5435) | offensive (0.5098) | co-defensive (0.8208) |
| bowl-winning (0.5818) | bowl-winning (0.6110) | bowl-winning (0.5886) | preseason (0.3921) | All-Pro (0.4310) | junior (0.3982) | preseason (0.4286) | ultra-athletic (0.7436) |
| preseason (0.5639) | undersized (0.5986) | preseason (0.5869) | drafted (0.3646) | All-American (0.4301) | preseason (0.3981) | All-American (0.4193) | N/A |

linewoman

| fastText | | | GloVe | | | | word2vec |
|---|---|---|---|---|---|---|---|
| Wikipedia | Gigawords | Wikipedia + Gigawords | Original | Wikipedia | Gigawords | Wikipedia + Gigawords | Original |
| N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |

lineperson

| fastText | | | GloVe | | | | word2vec |
|---|---|---|---|---|---|---|---|
| Wikipedia | Gigawords | Wikipedia + Gigawords | Original | Wikipedia | Gigawords | Wikipedia + Gigawords | Original |
| N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |

**Cosine Similarity in Adjectives for 'Woman'-Ending Profession Nouns in Word Embedding and their 'Man'- and 'Person'-Ending Counterparts**

chairwoman

| fastText | | | GloVe | | | | word2vec |
|---|---|---|---|---|---|---|---|
| Wikipedia | Gigawords | Wikipedia + Gigawords | Original | Wikipedia | Gigawords | Wikipedia + Gigawords | Original |
| emerita (0.4995) | emerita (0.5068) | emerita (0.5361) | democratic (0.2884) | non-executive (0.4312) | Democratic (0.3543) | non-executive (0.3402) | ex-officio (0.6693) |
| Inter-Parliamentary (0.4727) | nonprofit (0.4023) | non-executive (0.4615) | honorary (0.2848) | honorary (0.3361) | honorary (0.3381) | Governmental (0.3299) | ex-democratic (0.6667) |
| co-artistic (0.4636) | newly-created (0.3996) | co-artistic (0.4223) | supervisory (0.2773) | supervisory (0.3360) | Progressive (0.3018) | Democratic (0.3074) | ex-congressional (0.6652) |
| elect (0.4584) | honorary (0.3936) | Governmental (0.4220) | non-executive (0.2771) | Advisory (0.3232) | non-executive (0.2999) | supervisory (0.3037) | ex-republican (0.6451) |
| re-elected (0.4518) | Governmental (0.3933) | supervisory (0.4158) | emerita (0.2748) | elect (0.3181) | nonprofit (0.2997) | Republican (0.2919) | N/A |

chairman

| fastText | | | GloVe | | | | word2vec |
|---|---|---|---|---|---|---|---|
| Wikipedia | Gigawords | Wikipedia + Gigawords | Original | Wikipedia | Gigawords | Wikipedia + Gigawords | Original |
| non-executive (0.6899) | non-executive (0.6097) | non-executive (0.6208) | former (0.4092) | elect (0.5658) | Republican (0.4514) | Republican (0.4462) | non-executive (0.7882) |
| re-appointed (0.5751) | newly-created (0.4675) | supervisory (0.4740) | elected (0.4072) | non-executive (0.5409) | former (0.4375) | former (0.4387) | ex-officio (0.7548) |
| government-appointed (0.5267) | recently-appointed (0.4577) | newly-appointed (0.4464) | representative (0.3910) | representative (0.5316) | honorary (0.3967) | non-executive (0.4341) | re-appointed (0.6971) |
| supervisory (0.5212) | re-appointed (0.4530) | re-appointed (0.4441) | democratic (0.3810) | former (0.4964) | longtime (0.3902) | Judiciary (0.4053) | co-interim (0.6758) |
| ex-officio (0.5149) | honorary (0.4394) | Governmental (0.4383) | senior (0.3775) | interim (0.4751) | Judiciary (0.3865) | representative (0.4035) | ex-democratic (0.6732) |

chairperson

| fastText | | | GloVe | | | | word2vec |
|---|---|---|---|---|---|---|---|
| Wikipedia | Gigawords | Wikipedia + Gigawords | Original | Wikipedia | Gigawords | Wikipedia + Gigawords | Original |
| non-executive (0.5715) | inter-parliamentary (0.4873) | ex-officio (0.5487) | non-executive (0.4172) | elect (0.4561) | Parliamentary (0.3557) | non-executive (0.4676) | ex-officio (0.8495) |
| ex-officio (0.5682) | consultative (0.4816) | non-executive (0.5305) | elected (0.4048) | advisory (0.4068) | newly-elected (0.3488) | elect (0.4131) | chair-elect (0.7438) |
| government-appointed (0.5506) | newly-elected (0.4755) | consultative (0.5000) | honorary (0.3794) | ex-officio (0.3980) | All-China (0.3171) | ex-officio (0.4040) | non-professorial (0.6926) |
| elect (0.5299) | inter-ministerial (0.4642) | re-appointed (0.4947) | consultative (0.3400) | honorary (0.3843) | elect (0.3413) | honorary (0.3769) | advisory (0.6924) |
| state-appointed (0.5070) | ex-officio (0.4694) | newly-elected (0.4907) | ex-officio (0.3293) | interim (0.3913) | African (0.3332) | All-China (0.3658) | co-representative (0.6905) |

| congresswoman | | | | | | | |
|---|---|---|---|---|---|---|---|
| **fastText** | | | **GloVe** | | | | **word2vec** |
| **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| term-limited (0.5649) | congressional (0.5371) | congressional (0.5813) | incumbent (0.3702) | gubernatorial (0.3550) | senatorial (0.3331) | democratic (0.3322) | ex-republican (0.7529) |
| republican (0.5385) | republican (0.4980) | republican (0.5450) | republican (0.3590) | unseated (0.3545) | democratic (0.3205) | republican (0.3228) | congressional (0.7477) |
| vice-presidential (0.5359) | senatorial (0.4886) | senatorial (0.4949) | democratic (0.3578) | congressional (0.3258) | gubernatorial (0.2980) | congressional (0.3201) | republican (0.7423) |
| democratic (0.5358) | pro-choice (0.4840) | vice-presidential (0.4916) | representative (0.3276) | republican (0.3241) | congressional (0.2953) | incumbent (0.3152) | vice-presidential (0.7196) |
| incumbent (0.5259) | gubernatorial (0.4766) | democratic-held (0.4873) | gubernatorial (0.3105) | senatorial (0.3187) | pro-choice (0.2951) | gubernatorial (0.2994) | gubernatorial (0.7027) |

| congressman | | | | | | | |
|---|---|---|---|---|---|---|---|
| **fastText** | | | **GloVe** | | | | **word2vec** |
| **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| congressional (0.7957) | congressional (0.6531) | congressional (0.7101) | republican (0.6145) | congressional (0.6442) | congressional (0.6049) | congressional (0.6557) | republican (0.8268) |
| republican (0.6520) | republican (0.5668) | republican (0.6270) | congressional (0.5593) | incumbent (0.6385) | republican (0.5841) | republican (0.6092) | ex-republican (0.7902) |
| republican-controlled (0.5985) | senatorial (0.5302) | republican-leaning (0.5599) | representative (0.5022) | republican (0.6208) | democratic (0.5003) | democratic (0.5305) | vice-presidential (0.7800) |
| democratic-controlled (0.5942) | republican-held (0.5101) | democratic-held (0.5520) | democratic (0.5012) | democratic (0.5292) | incumbent (0.4639) | incumbent (0.5011) | gubernatorial (0.7687) |
| democratic-leaning (0.5919) | gubernatorial (0.5123) | gubernatorial (0.5386) | incumbent (0.4804) | gubernatorial (0.4907) | former (0.4519) | gubernatorial (0.4585 | congressional (0.7848) |

| congressperson | | | | | | | |
|---|---|---|---|---|---|---|---|
| **fastText** | | | **GloVe** | | | | **word2vec** |
| **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| congressional (0.7442) | N/A | congressional (0.6921) | N/A | reform-minded (0.3143 | N/A | unchurched (0.3286) | libertarian (0.7162) |
| republican-controlled (0.5731) | N/A | republican (0.5486) | N/A | unchurched (0.2954) | N/A | pro-obama (0.3060) | ultra-liberals (0.7141) |
| democratic-controlled (0.5524) | N/A | non-republican (0.5312) | N/A | sector-specific (0.2914) | N/A | community-minded (0.3052) | non-republican (0.7128) |
| Senatorial (0.5438) | N/A | republican-leaning (0.5305) | N/A | Incumbent (0.2846) | N/A | self-admitted (0.3037) | republican (0.7102) |
| democratic-leaning (0.5421) | N/A | republican-dominated (0.5247) | N/A | metalinguistic (0.2652) | N/A | bristol-based (0.2998) | pro-gun (0.7042) |

| businesswoman | | | | | | | |
|---|---|---|---|---|---|---|---|
| **fastText** | | | **GloVe** | | | | **word2vec** |
| **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| née (0.5690) | self-made (0.5044) | née (0.5187) | resourceful (0.3898) | aspiring (0.4457) | self-made (0.4159) | self-made (0.4282) | self-made (0.4282) |
| nigerian-born (0.5221) | well-connected (0.4969) | self-made (0.4958) | strong-willed (0.3657) | british-born (0.4369) | wealthy (0.4034) | aspiring (0.4213) | well-connected (0.4246) |
| american-born (0.5204) | auburn-haired (0.4841) | well-connected (0.4926) | british-born (0.3593) | canadian-born (0.3884) | well-connected (0.3988) | well-connected (0.4246) | aspiring (0.4213) |
| british-born (0.5176) | wealthy (0.4744) | malaysian-born (0.4925) | feisty (0.3580) | née (0.3875) | prominent (0.3878) | wealthy (0.3994) | née (0.3845) |
| english-born (0.5062) | vivacious (0.4684) | aspiring (0.4828) | german-born (0.3428) | russian-born (0.3854) | beijing-born (0.3674) | née (0.3845) | prominent (0.3684) |

| businessman | | | | | | | |
|---|---|---|---|---|---|---|---|
| **fastText** | | | **GloVe** | | | | **word2vec** |
| **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| wealthy (0.6595) | wealthy (0.5765) | wealthy (0.5977) | wealthy (0.5572) | wealthy (0.6975) | wealthy (0.6378) | wealthy (0.6574) | octogenarian (0.7143) |
| scottish-born (0.5479) | well-connected (0.5502) | self-made (0.5606) | born (0.4757) | prominent (0.6157) | prominent (0.5899) | prominent (0.5814) | wealthy (0.6812) |
| self-made (0.5440) | self-made (0.5349) | well-connected (0.5543) | prominent (0.4637) | well-known (0.5349) | well-known (0.4826) | well-known (0.4836) | septuagenarian (0.6740) |
| english-born (0.5329) | american-educated (0.4858) | well-to-do (0.5140) | former (0.3909) | former (0.4920) | taiwanese (0.4514) | self-made (0.4650) | canadian- (0.6740) |
| american-born (0.5284) | prominent (0.4843) | american-educated (0.5011) | convicted (0.3867) | renowned (0.4828) | well-connected (0.4497) | retired (0.4379) | ukrainian- (0.6496) |

| businessperson | | | | | | | |
|---|---|---|---|---|---|---|---|
| **fastText** | | | **GloVe** | | | | **word2vec** |
| **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| norwegian-born (0.6129) | business-minded (0.5346) | norwegian-american (0.5443) | german-born (0.4653) | british-born (0.4223) | philippine-born (0.3532) | norwegian-american (0.3998) | canadian- (0.6576) |
| danish-born (0.5827) | well-educated (0.5189) | english-born (0.5352) | british-born (0.4400) | german-born (0.4096) | home-based (0.3379) | english-born (0.3898) | born (0.6434) |
| norwegian-american (0.5682) | entrepreneurial (0.4997) | scottish-born (0.5170) | russian-born (0.4107) | norwegian (0.4031) | unselfish (0.3371) | norwegian (0.3880) | N/A |
| indian-born (0.5624) | business-related (0.4990) | canadian-american (0.5140) | irish-born (0.4107) | scottish-born (0.3992) | still-young (0.3371) | hungarian-american (0.3691) | N/A |
| scottish-born (0.5611) | well-connected (0.4855) | english-american (0.5010) | scottish-born (0.3974) | irish-born (0.3898) | unadventurous (0.3353) | canadian-american (0.3659) | N/A |

councilwoman

| fastText | | | GloVe | | | | word2vec |
|---|---|---|---|---|---|---|---|
| **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| mayoral (0.6472) | mayoral (0.5013) | mayoral (0.5608) | mayoral (0.2698) | openly-gay (0.3184) | mayoral (0.2537) | mayoral (0.3331) | mayoral (0.7357) |
| term-limited (0.5526) | term-limited (0.4217) | term-limited (0.4758) | nymphomaniac (0.2383) | mayoral (0.2935) | non-mormon (0.2453) | liveable (0.242) | N/A |
| democratic-leaning (0.4828) | Libertarian (0.4168) | crime-plagued (0.4374) | shrewish (0.2290) | directly-elected (0.2872) | write-in (0.2365) | non-mormon (0.2455) | N/A |
| incumbent (0.4788) | anti-gang (0.4074) | pro-development (0.4293) | rhotic (0.2228) | county-controlled (0.2756) | ultraconservative (0.2259) | arab-jewish (0.2421) | N/A |
| re-elected (0.4672) | crime-plagued (0.4032) | anti-gang (0.4117) | rewritable (0.2164) | incumbent (0.2738) | ethnically-mixed (0.2223) | anti-gang (0.2404) | N/A |

councilman

| fastText | | | GloVe | | | | word2vec |
|---|---|---|---|---|---|---|---|
| **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| mayoral (0.7028) | mayoral (0.5798) | mayoral (0.6321) | mayoral (0.4098) | mayoral (0.5096) | mayoral (0.4156) | mayoral (0.4511) | mayoral (0.7651) |
| aldermanic (0.6264) | term-limited (0.4969) | aldermanic (0.4943) | incumbent (0.3676) | incumbent (0.4504) | Albanese (0.3599) | incumbent (0.3486) | aldermanic (0.7186) |
| reelect (0.5846) | Libertarian (0.4306) | republican (0.4354) | albanese (0.3666) | re-elected (0.4264) | municipal (0.2582) | re-elected (0.3283) | N/A |
| term-limited (0.5548) | anti-gang (0.4212) | gubernatorial (0.4225) | re-elected (0.3292) | republican (0.3618) | Socialist (0.2555) | municipal (0.3135) | N/A |
| elect (0.5538) | gang-infested (0.4149) | re-elected (0.4225) | elected (0.3158) | municipal (0.3584) | drive-by (0.2538) | Albanese (0.3022) | N/A |

councilperson

| fastText | | | GloVe | | | | word2vec |
|---|---|---|---|---|---|---|---|
| **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** | **Wikipedia** | **Gigawords** | **Wikipedia + Gigawords** | **Original** |
| directly-elected (0.6044) | N/A | N/A | N/A | ex-officio (0.3366) | N/A | N/A | N/A |
| elect (0.5576) | N/A | N/A | N/A | subdivisional (0.3130) | N/A | N/A | N/A |
| N/A | N/A | N/A | N/A | Frightful (0.3096) | N/A | N/A | N/A |
| N/A | N/A | N/A | N/A | non-consecutive (0.2976) | N/A | N/A | N/A |
| N/A | N/A | N/A | N/A | political-military (0.2792) | N/A | N/A | N/A |

## Semantic Orientation Quartile Distribution for Extreme Masculine Professions in Labour Statistics from Spain

| | fastText | GloVe | word2vec |
|---|---|---|---|
| *Constructor* | 0,2977 | 0,3897 | 0,3263 |
| *Carpintero* | 0,2338 | 0,3380 | 0,4356 |
| *Ingeniero* | 0,2507 | 0,3821 | 0,2046 |
| *Pescador* | 0,1632 | 0,3195 | 0,3895 |
| *Zapatero* | 0,0451 | -0,0039 | 0,2386 |
| *Ebanista* | 0,1055 | 0,3600 | 0,2764 |
| *Reparador* | 0,2730 | 0,2429 | 0,1522 |
| *Agricultor* | 0,2292 | 0,3850 | 0,3021 |
| *Ganadero* | 0,1950 | 0,3352 | 0,2390 |
| *Informático* | 0,1250 | 0,3207 | 0,1318 |
| *Low Bias Limit (fem)* | 0 — –0,0315 | 0 — –0,0565 | 0 — –0,0281 |
| *Medium Bias Limit (fem)* | –0,0316 — – 0,0693 | –0,0566 — – 0,1169 | –0,0282 — – 0,0625 |
| *High Bias Limit (fem)* | –0,0694 — – 0,1265 | –0,1170 — – 0,1989 | –0,0626 — – 0,1161 |
| *Extremely High Bias Limit (fem)* | Less than – 0,1266 | Less than – 0,1990 | Less than – 0,1162 |
| *Low Bias Limit (mas)* | 0 – 0,0359 | 0 — 0,0390 | 0 — 0,0299 |
| *Medium Bias Limit (mas)* | 0,0360 – 0,0751 | 0,0391 — 0,0849 | 0,0300 — 0,0631 |
| *High bias Limit (mas)* | 0,0752 – 0,1267 | 0,0850 — 0,1485 | 0,0632 — 0,1079 |
| *Extremely High Bias Limit (mas)* | More than 0,1268 | More than 0,1486 | More than 0,1080 |

| | fastText | GloVe | word2vec |
|---|---|---|---|
| *Constructora* | -0,1767 | -0,2271 | -0,2009 |
| *Carpintera* | -0,0408 | -0,2147 | -0,0222 |
| *Ingeniera* | -0,5628 | -0,4688 | -0,5941 |
| *Pescadora* | -0,2775 | -0,1649 | -0,0388 |
| *Zapatera* | -0,1926 | -0,1177 | -0,3429 |
| *Ebanista* | = | = | = |
| *Reparadora* | -0,2603 | -0,2674 | -0,2035 |
| *Agricultora* | -0,3903 | -0,2049 | -0,2589 |
| *Ganadera* | 0,0058 | -0,0588 | -0,0322 |
| *Informática* | -0,0322 | -0,0480 | -0,0695 |
| *Low Bias Limit (fem)* | 0 — –0,0315 | 0 — –0,0565 | 0 — –0,0281 |
| *Medium Bias Limit (fem)* | –0,0316 — – 0,0693 | –0,0566 — – 0,1169 | –0,0282 — – 0,0625 |
| *High Bias Limit (fem)* | –0,0694 — – 0,1265 | –0,1170 — – 0,1989 | –0,0626 — – 0,1161 |
| *Extremely High Bias Limit (fem)* | Less than – 0,1266 | Less than – 0,1990 | Less than – 0,1162 |
| *Low Bias Limit (mas)* | 0 – 0,0359 | 0 — 0,0390 | 0 — 0,0299 |
| *Medium Bias Limit (mas)* | 0,0360 – 0,0751 | 0,0391 — 0,0849 | 0,0300 — 0,0631 |
| *High bias Limit (mas)* | 0,0752 – 0,1267 | 0,0850 — 0,1485 | 0,0632 — 0,1079 |
| *Extremely High Bias Limit (mas)* | More than 0,1268 | More than 0,1486 | More than 0,1080 |

## Semantic Orientation Quartile Distribution for Extreme Feminine Professions in Labour Statistics from Spain

| | fastText | GloVe | word2vec |
|---|---|---|---|
| *Asistente* | -0,0107 | -0,0246 | -0,0091 |
| *Costurera* | -0,4618 | -0,5247 | -0,4225 |
| *Bibliotecaria* | -0,5006 | -0,3468 | -0,4032 |
| *Archivista* | 0,0293 | 0,0732 | 0,0620 |
| *Educadora* | -0,6193 | -0,5541 | -0,5756 |
| *Veterinaria* | -0,0430 | -0,0243 | -0,0540 |
| *Jardinera* | -0,3344 | -0,4185 | -0,2778 |
| *Auxiliar* | 0,0653 | 0,0941 | 0,0834 |
| *Gestora* | -0,3234 | -0,3189 | -0,3033 |
| *Investigadora* | -0,4672 | -0,5045 | -0,5064 |
| *Low Bias Limit (fem)* | 0 — –0,0315 | 0 — –0,0565 | 0 — –0,0281 |
| *Medium Bias Limit (fem)* | –0,0316 — – 0,0693 | –0,0566 — – 0,1169 | –0,0282 — – 0,0625 |
| *High Bias Limit (fem)* | –0,0694 — – 0,1265 | –0,1170 — – 0,1989 | –0,0626 — – 0,1161 |
| *Extremely High Bias Limit (fem)* | Less than – 0,1266 | Less than – 0,1990 | Less than – 0,1162 |
| *Low Bias Limit (mas)* | 0 – 0,0359 | 0 — 0,0390 | 0 — 0,0299 |
| *Medium Bias Limit (mas)* | 0,0360 – 0,0751 | 0,0391 — 0,0849 | 0,0300 — 0,0631 |
| *High bias Limit (mas)* | 0,0752 – 0,1267 | 0,0850 — 0,1485 | 0,0632 — 0,1079 |
| *Extremely High Bias Limit (mas)* | More than 0,1268 | More than 0,1486 | More than 0,1080 |

| | fastText | GloVe | word2vec |
|---|---|---|---|
| *Asistente* | = | = | = |
| *Costurero* | -0,1467 | -0,2676 | -0,0549 |
| *Bibliotecario* | 0,1503 | 0,2561 | 0,2344 |
| *Archivista* | = | = | = |
| *Educador* | 0,3099 | 0,5030 | 0,3381 |
| *Veterinario* | 0,1407 | 0,2695 | 0,2466 |
| *Jardinero* | 0,2207 | 0,3501 | 0,2490 |
| *Auxiliar* | = | = | = |
| *Gestor* | 0,3278 | 0,4036 | 0,3211 |
| *Investigador* | 0,2690 | 0,4772 | 0,2354 |
| *Low Bias Limit (fem)* | 0 — –0,0315 | 0 — –0,0565 | 0 — –0,0281 |
| *Medium Bias Limit (fem)* | –0,0316 — – 0,0693 | –0,0566 — – 0,1169 | –0,0282 — – 0,0625 |
| *High Bias Limit (fem)* | –0,0694 — – 0,1265 | –0,1170 — – 0,1989 | –0,0626 — – 0,1161 |
| *Extremely High Bias Limit (fem)* | Less than – 0,1266 | Less than – 0,1990 | Less than – 0,1162 |
| *Low Bias Limit (mas)* | 0 – 0,0359 | 0 — 0,0390 | 0 — 0,0299 |
| *Medium Bias Limit (mas)* | 0,0360 – 0,0751 | 0,0391 — 0,0849 | 0,0300 — 0,0631 |
| *High bias Limit (mas)* | 0,0752 – 0,1267 | 0,0850 — 0,1485 | 0,0632 — 0,1079 |
| *Extremely High Bias Limit (mas)* | More than 0,1268 | More than 0,1486 | More than 0,1080 |

## Semantic Orientation Quartile Distribution for Gender-Neutral Professions in Labour Statistics from Spain

| | fastText | GloVe | word2vec |
|---|---|---|---|
| *Financiero* | 0,0274 | 0,1428 | 0,0009 |
| *Comerciante* | 0,1735 | 0,2499 | 0,2363 |
| *Científico* | 0,1748 | 0,3075 | 0,3699 |
| *Farmaceuta* | -0,2066 | -0,0077 | -0,1342 |
| *Editor* | 0,1756 | 0,2601 | 0,1014 |
| *Presentador* | 0,1428 | 0,3067 | 0,0303 |
| *Publicista* | -0,0344 | 0,0693 | -0,0425 |
| *Funcionario* | 0,1666 | 0,2211 | 0,1349 |
| *Artista* | -0,0417 | -0,0458 | -0,0101 |
| *Low Bias Limit (fem)* | 0 — –0,0315 | 0 — –0,0565 | 0 — –0,0281 |
| *Medium Bias Limit (fem)* | –0,0316 — –0,0693 | –0,0566 — –0,1169 | –0,0282 — –0,0625 |
| *High Bias Limit (fem)* | –0,0694 — –0,1265 | –0,1170 — –0,1989 | –0,0626 — –0,1161 |
| *Extremely High Bias Limit (fem)* | Less than –0,1266 | Less than –0,1990 | Less than –0,1162 |
| *Low Bias Limit (mas)* | 0 – 0,0359 | 0 — 0,0390 | 0 — 0,0299 |
| *Medium Bias Limit (mas)* | 0,0360 – 0,0751 | 0,0391 — 0,0849 | 0,0300 — 0,0631 |
| *High bias Limit (mas)* | 0,0752 – 0,1267 | 0,0850 — 0,1485 | 0,0632 — 0,1079 |
| *Extremely High Bias Limit (mas)* | More than 0,1268 | More than 0,1486 | More than 0,1080 |

| | fastText | GloVe | word2vec |
|---|---|---|---|
| *Financiera* | -0,0837 | -0,1145 | -0,1051 |
| *Comerciante* | = | = | = |
| *Científica* | -0,0414 | -0,1066 | -0,0198 |
| *Farmaceuta* | = | = | = |
| *Editora* | -0,5208 | -0,5945 | -0,6043 |
| *Presentadora* | -0,6383 | -0,7147 | -0,7037 |
| *Publicista* | = | = | = |
| *Funcionaria* | -0,5811 | -0,6330 | -0,6183 |
| *Artista* | = | = | = |
| *Low Bias Limit (fem)* | 0 — –0,0315 | 0 — –0,0565 | 0 — –0,0281 |
| *Medium Bias Limit (fem)* | –0,0316 — –0,0693 | –0,0566 — –0,1169 | –0,0282 — –0,0625 |
| *High Bias Limit (fem)* | –0,0694 — –0,1265 | –0,1170 — –0,1989 | –0,0626 — –0,1161 |
| *Extremely High Bias Limit (fem)* | Less than –0,1266 | Less than –0,1990 | Less than –0,1162 |
| *Low Bias Limit (mas)* | 0 – 0,0359 | 0 — 0,0390 | 0 — 0,0299 |
| *Medium Bias Limit (mas)* | 0,0360 – 0,0751 | 0,0391 — 0,0849 | 0,0300 — 0,0631 |
| *High bias Limit (mas)* | 0,0752 – 0,1267 | 0,0850 — 0,1485 | 0,0632 — 0,1079 |
| *Extremely High Bias Limit (mas)* | More than 0,1268 | More than 0,1486 | More than 0,1080 |

**Cosine Similarity in Adjectives for Top Extreme Masculine Profession Nouns and Their Female Counterparts in Labour Statistics from Spain**

| constructor | | | ebanista | | |
|---|---|---|---|---|---|
| **fastText** | **GloVe** | **word2vec** | **fastText** | **GloVe** | **word2vec** |
| constructor (1.0000) | constructor (1.0000) | constructor (1.0000) | constructor (0.5658) | constructor (0.3992) | calcográfico (0.5554) |
| constructora (0.6572) | promotor (0.4946) | constructora (0.6043) | ilustrador (0.5592) | reputado (0.3804) | autodidacto (0.5459) |
| constructores (0.6263) | propietario (0.4765) | constructores (0.5734) | autodidacta (0.5351) | afamado (0.3518) | litográfico (0.5439) |
| edificador (0.5510) | constructora (0.4645) | comanditario (0.5342) | proyectista (0.5311) | acaudalado (0.3488) | ilustrador (0.5304) |
| propietario (0.53538) | inventor (0.4591) | propietario (0.5288) | trinchador (0.5220) | proyectista (0.3317) | constructor (0.5249) |
| constructora | | | | | |
| **fastText** | **GloVe** | **word2vec** | | | |
| constructora (1.0000) | constructora (1.0000) | constructora (1.0000) | | | |
| constructoras (0.7450) | concesionaria (0.5853) | concesionaria (0.7105) | | | |
| concesionaria (0.7218) | promotora (0.5469) | constructoras (0.7079) | | | |
| constructor (0.6572) | ferrovial (0.5353) | constructor (0.6043) | | | |
| ferrovial (0.6329) | constructoras (0.4945) | opada (0.6011) | | | |

**Cosine Similarity in Adjectives for Top Extreme Feminine Profession Nouns and Their Male Counterparts in Labour Statistics from Spain**

| asistente | | | costurera | | | bibliotecaria | | |
|---|---|---|---|---|---|---|---|---|
| **fastText** | **GloVe** | **word2vec** | **fastText** | **GloVe** | **word2vec** | **fastText** | **GloVe** | **word2vec** |
| asesor (0.5458) | nombrado (0.5826) | asesor (0.6215) | pinturera (0.5656) | humilde (0.4190) | humilde (0.5911) | traductora (0.6121) | traductora (0.3599) | traductora (0.5890) |
| suplente (0.5323) | asesor (0.5490) | Criminólogo (0.5395) | huérfana (0.5599) | vendedora (0.3989) | soltera (0.5710) | bibliotecológico (0.6106) | investigadora (0.3354) | bibliófila (0.5825) |
| nombrado (0.5169) | investigador (0.5400) | Fisiológico (0.5333) | solterona (0.5454) | huérfana (0.3794) | huérfana (0.5702) | bibliófila (0.5882) | simpática (0.3223) | divulgadora (0.5782) |
| supervisor (0.5113) | designado (0.5025) | cardiotorácico (0.5330) | humilde (0.5449) | ilustradora (0.3609) | afanada (0.5585) | educadora (0.5397) | veinteañera (0.3201) | emérita (0.5761) |
| emérito (0.4929) | instructor (0.4617) | Estimuladora (0.5317) | soltera (0.5398) | devota (0.3582) | carnicero (0.5503) | bibliográfica (0.5356) | tímida (0.3190) | estudiosa (0.5618) |
| asistenta | | | costurero | | | bibliotecario | | |
| **fastText** | **GloVe** | **word2vec** | **fastText** | **GloVe** | **word2vec** | **fastText** | **GloVe** | **word2vec** |
| cuidadora (0.7073) | cuidadora (0.3504) | cuidadora (0.6453) | estampador (0.5592) | saltarín (0.2570) | rameado (0.6453) | bibliófilo (0.6229) | nombrado (0.4769) | traductor (0.6144) |
| anciana (0.6363) | educadora (0.3350) | anciana (0.6320) | confeccionista (0.5530) | castellanoleonés (0.2518) | almidonada (0.6284) | bibliotecológico (0.6048) | traductor (0.4683) | bibliófilo (0.6125) |
| maltratadora (0.6287) | anciana (0.2922) | treintañera (0.6203) | confeccionador (0.5434) | barrigona (0.2515) | trinchador (0.6220) | traductor (0.6029) | investigador (0.4327) | bibliográfico (0.5424) |
| treintañera (0.5883) | pía (0.2806) | cuarentona (0.5975) | lagarterano (0.5245) | fosfatado (0.2504) | descotada (0.6193) | bibliográfico (0.5793) | emérito (0.4204) | educador (0.5405) |
| drogadicta (0.5774) | censitaria (0.2706) | realquilada (0.5824) | lagarterana (0.5231) | mineralógico (0.2435) | laqueado (0.6190) | educador (0.5585) | estudioso (0.3890) | Trilingüe (0.5325) |

## Semantic Orientation for Extreme Masculine Professions in Word Embeddings in Spanish

| fastText | GloVe | word2vec |
|---|---|---|
| acopiador (0.4578) | cuidador (0.5702) | sabio (0.4977) |
| malhechor (0.4300) | discípulo (0.5688) | capuchino (0.4937) |
| timbalero (0.4193) | filósofo (0.5515) | excavador (0.4771) |
| oteador (0.4125) | educador (0.5030) | carpintero (0.4356) |
| discípulo (0.3953) | biólogo (0.4859) | pensador (0.4205) |
| pensador (0.5920) | confitero (0.4777) | comentador (0.4164) |
| thaumaturgo (0.3918) | investigador (0.4772) | artesano (0.4126) |
| zelador (0.3876) | antropólogo (0.4749) | cuidador (0.4042) |
| baqueano (0.3865) | sociólogo (0.4716) | discípulo (0.4023) |
| techador (0.3750) | franciscano (0.7710) | cantor (0.4015) |

## Semantic Orientation for Extreme Feminine Professions in Word Embeddings in Spanish

| fastText | GloVe | word2vec |
|---|---|---|
| empresaria (-0.7447) | poetisa (-0.7980) | empresaria (-0.7608) |
| podóloga (-0.7031) | fotógrafa (-0.7938) | politóloga (-0.7434) |
| sinóloga (-0.6992) | ventrílocua (-0.7811) | jefa (-0.7343) |
| directora (-0.6896) | empresaria (-0.7477) | directora (-0.7338) |
| ginecóloga (-0.6876) | rejoneadora (-0.7404) | escritora (-0.7307) |
| exactriz (-0.6781) | actriz (-0.7301) | abogada (-0.7195) |
| aviadora (-0.6694) | presentadora (-0.7147) | presidenta (-0.7068) |
| escritora (-0.6676) | embajadora (-0.7123) | presentadora (-0.7037) |
| repositora (-0.6618) | bailarina (-0.7072) | actriz (-0.7016) |
| abogada (-0.6571) | camarera (-0.7059) | locutora (-0.6965) |

**Cosine Similarity in Adjectives for Top Extreme Masculine Profession Nouns and Their Female Counterparts in Word Embedding**

| discípulo | | | cuidador | | | pensador | | |
|---|---|---|---|---|---|---|---|---|
| **fastText** | **GloVe** | **word2vec** | **fastText** | **GloVe** | **word2vec** | **fastText** | **GloVe** | **word2vec** |
| seguidor (0.6459) | seguidor (0.5730) | seguidor (0.6689) | discapacitado (0.5523) | discapacitado (0.4542) | discapacitado (0.5867) | librepensador (0.7066) | estudioso (0.5799) | teorizador (0.6953) |
| aventajado (0.6407) | aventajado (0.5370) | continuador (0.6019) | asustador (0.5312) | adulto (0.4489) | paseador (0.5371) | humanista (0.6592) | humanista (0.5415) | librepensador (0.6936) |
| admirador (0.6171) | admirador (0.5322) | estudioso (0.5910) | caminador (0.5172) | adoptivo (0.3789) | extrafamiliares (0.5267) | estudioso (0.6561) | eminente (0.5129) | enciclopedista (0.6838) |
| estudioso (0.6017) | estudioso (0.4774) | aventajado (0.5877) | trotador (0.5091) | anciano (0.3513) | desavenido (0.5245) | filosófico (0.6450) | educador (0.4836) | humanista (0.6828) |
| continuador (0.5297) | continuador (0.4624) | eleática (0.5825) | intimidador (0.5084) | parental (0.3485) | sordomudo (0.5192) | enciclopedista (0.6385) | traductor (0.4697) | pragmatista (0.6787) |
| **discípula** | | | **cuidadora** | | | **pensadora** | | |
| **fastText** | **GloVe** | **word2vec** | **fastText** | **GloVe** | **word2vec** | **fastText** | **GloVe** | **word2vec** |
| admiradora (0.5702) | aventajada (0.3773) | estudiosa (0.5650) | maltratadora (0.6768) | anciana (0.3565) | anciana (0.6480) | cansadora (0.7463) | ilustradora (0.3454) | feminista (0.7063) |
| estudiosa (0.5672) | predilecta (0.3713) | coterránea (0.5621) | anciana (0.6556) | jubilada (0.3168) | heroinómana (0.6015) | librepensadora (0.7274) | aventurera (0.3434) | librepensadora (0.6757) |
| aventajada (0.5516) | psicoanalista (0.3623) | devota (0.5574) | celadora (0.6454) | adorada (0.2976) | preadoptivos (0.5825) | prensadora (0.6610) | desinhibida (0.3299) | agitadora (0.6249) |
| seguidora (0.5450) | estudiosa (0.3291) | virtuosa (0.5518) | intimidadora (0.6152) | repartidor (0.2975) | toxicómana (0.5787) | dispensadora (0.6543) | riquísima (0.3239) | objetivista (0.6173) |
| adoradora (0.5217) | admiradora (0.3268) | ventrílocua (0.5501) | maltratada (0.5990) | catalizadora (0.2952) | huerfanas (0.5765) | feminista (0.6328) | unitaria (0.3208) | sufragista (0.6102) |

**Cosine Similarity in Adjectives for Top Extreme Feminine Profession Nouns and Their Masculine Counterparts in Word Embeddings**

| empresaria | | | abogada | | | actriz | | |
|---|---|---|---|---|---|---|---|---|
| **fastText** | **GloVe** | **word2vec** | **fastText** | **GloVe** | **word2vec** | **fastText** | **GloVe** | **word2vec** |
| comunicadora (0.6637) | educadora (0.5105) | comunicadora (0.6607) | penalista (0.6393) | experta (0.4925) | matrimonialista (0.6758) | porno (0.6291) | nacida (0.6309) | ventrílocua (0.6315) |
| educadora (0.6624) | fundadora (0.4742) | educadora (0.6295) | abogador (0.6001) | investigadora (0.4769) | criminóloga (0.6445) | narradora (0.6259) | fallecida (0.5515) | declamadora (0.6308) |
| emprendedora (0.6463) | nacida (0.4700) | declamadora (0.6123) | criminóloga (0.5788) | penalista (0.4601) | penalista (0.6195) | actoral (0.6211) | joven (0.5398) | porno (0.6207) |
| cultora (0.6014) | conductora (0.4638) | cucuteña (0.6075) | educadora (0.5733) | matrimonialista (0.4529) | educadora (0.5969) | jugadora (0.6016) | porno (0.5380) | bulímica (0.6164) |
| empresarial (0.5853) | empresarial (0.4486) | recitadora (0.5995) | comunicadora (0.5678) | fundadora (0.4458) | comunicadora (0.5710) | guapísima (0.5988) | ganadora (0.5291) | narradora (0.6134) |

| empresario | | | abogado | | | actor | | |
|---|---|---|---|---|---|---|---|---|
| **fastText** | **GloVe** | **word2vec** | **fastText** | **GloVe** | **word2vec** | **fastText** | **GloVe** | **word2vec** |
| multimillonario (0.6777) | propietario (0.6749) | propietario (0.6857) | penalista (0.7254) | investigador (0.5916) | penalista (0.7629) | cinematográfico (0.6897) | cinematográfico (0.5815) | cinematográfico (0.6865) |
| millonario (0.6624) | promotor (0.5961) | multimillonario (0.6548) | litigante (0.5690) | experto (0.5772) | litigante (0.6967) | actoral (0.6572) | veterano (0.5789) | actoral (0.6188) |
| propietario (0.6565) | multimillonario (0.5816) | millonario (0.6418) | querellante (0.5786) | penalista (0.5596) | administrativista (0.6227) | teatral (0.6361) | teatral (0.5564) | pornografico (0.6166) |
| inversionista (0.6212) | millonario (0.5792) | acaudalado (0.6228) | acusador (0.5486) | asesor (0.5342) | criminólogo (0.6182) | cinematográfica (0.5998) | joven (0.5518) | teatral (0.6126) |
| acaudalado (0.6114) | constructor (0.5780) | adinerado (0.6012) | criminólogo (0.5425) | imputado (0.5244) | matrimonialista (0.6135) | histriónico (0.5757) | destacado (0.5415) | guapísimo (0.6162) |

| directora | | | escritora | | | presentadora | | |
|---|---|---|---|---|---|---|---|---|
| **fastText** | **GloVe** | **word2vec** | **fastText** | **GloVe** | **word2vec** | **fastText** | **GloVe** | **word2vec** |
| coordinadora (0.6364) | coordinadora (0.6433) | coordinadora (0.7618) | narradora (0.7647) | traductora (0.6080) | traductora (0.7626) | reportera (0.7538) | conductora (0.6573) | reportera (0.7451) |
| asesora (0.6134) | investigadora (0.5746) | fundadora (0.6127) | traductora (0.7618) | nacida (0.5840) | narradora (0.7576) | conductora (0.7261) | reportera (0.6469) | conductora (0.7170) |
| fundadora (0.6119) | asesora (0.5629) | delegada (0.6119) | ilustradora (0.7243) | educadora (0.5341) | ilustradora (0.7286) | tertuliana (0.6683) | veterana (0.5181) | tertuliana (0.7026) |
| directoral (0.6062) | fundadora (0.5481) | cofundadora (0.5955) | educadora (0.6839) | literaria (0.5246) | educadora (0.6799) | comunicadora (0.6406) | televisiva (0.5115) | comunicadora (0.6910) |
| delegada (0.6042) | delegada (0.5301) | investigadora (0.5861) | nacida (0.6457) | feminista (0.5246) | nacida (0.6440) | televisivo (0.6141) | televisivo (0.4917) | televisivo (0.6380) |

| director | | | escritor | | | presentador | | |
|---|---|---|---|---|---|---|---|---|
| **fastText** | **GloVe** | **word2vec** | **fastText** | **GloVe** | **word2vec** | **fastText** | **GloVe** | **word2vec** |
| directoral (0.6271) | asesor (0.6399) | coordinador (0.6455) | narrador (0.7272) | literario (0.6385) | literario (0.7149) | televisivo (0.7217) | televisivo (0.6572) | televisivo (0.7342) |
| asesor (0.6047) | delegado (0.5950) | asesor (0.6341) | traductor (0.7142) | traductor (0.6375) | traductor (0.7090) | radiofónico (0.7128) | reportero (0.5894) | radiofónico (0.6979) |
| delegado (0.5952) | coordinador (0.5941) | supervisor (0.6104) | ilustrador (0.6864) | narrador (0.6339) | ilustrador (0.6757) | reportero (0.6885) | radiofónico (0.5635) | comunicador (0.6686) |
| supervisor (0.5577) | investigador (0.5926) | delegado (0.5985) | literario (0.6795) | creador (0.5942) | estudioso (0.6554) | tertuliana (0.6333) | comunicador (0.5546) | reportero (0.6380) |
| coordinador (0.5518) | nombrado (0.5667) | investigador (0.5710) | estudioso (0.6690) | célebre (0.5866) | divulgador (0.6444) | comunicador (0.6207) | veterano (0.5212) | tertuliano (0.6267) |