




Article

Decision Algorithm for the Automatic Determination of the Use of Non-Inclusive Terms in Academic Texts

Pedro Orgeira-Crespo ^{1,*}, Carla Míguez-Álvarez ², Miguel Cuevas-Alonso ² and María Isabel Doval-Ruiz ³

¹ Aerospace Area, Department of Mechanical Engineering, Heat Engines and Machines, and Fluids, Aerospace Engineering School, University of Vigo, Campus Orense, 32004 Orense, Spain

² Language Variation and Textual Categorization (LVTC), Philology and Translation School, University of Vigo, 36310 Vigo, Spain; camiguez@uvigo.es (C.M.-Á.); miguel.cuevas@uvigo.es (M.C.-A.)

³ Faculty of Educational Sciences, University of Vigo, Campus Lagoas Marcosende, 36310 Vigo, Spain; mdoval@uvigo.es

* Correspondence: porgeira@uvigo.es

Received: 9 April 2020; Accepted: 3 August 2020; Published: 6 August 2020



Abstract: The use of inclusive language, among many other gender equality initiatives in society, has garnered great attention in recent years. Gender equality offices in universities and public administration cannot cope with the task of manually checking the use of non-inclusive language in the documentation that those institutions generate. In this research, an automated solution for the detection of non-inclusive uses of the Spanish language in doctoral theses generated in Spanish universities is introduced using machine learning techniques. A large dataset has been used to train, validate, and analyze the use of inclusive language; the result is an algorithm that detects, within any Spanish text document, non-inclusive uses of the language with error, false positive, and false negative ratios slightly over 10%, and precision, recall, and F-measure percentages over 86%. Results also show the evolution with time of the ratio of non-inclusive usages per document, having a pronounced reduction in the last years under study.

Keywords: inclusive language; Spanish language; natural language processing; classification algorithm; machine learning

1. Introduction

This research focuses on the problem of identifying without human supervision whether the use of a term in its context inside a long portion of text is considered non-inclusive in the modern Spanish language. A term itself does not have the characteristic of being inclusive or not: It is how it is utilized in context that determines whether the use of the term is inclusive or not. For instance, the word “*profesor*” (teacher) may be used in a context where it is not considered non-inclusive: “... *los profesores Ricardo Marín y Enrique Mandado, creadores de los grupos de ...*” (professors Ricardo and Enrique, creators of the groups ...); on the other hand, it may be used in a different context where it is considered non-inclusive: “... *en el aula, los profesores debería tener autoridad para que haya respeto ...*” (in the classroom, teachers should have authority for respect to be ...), having one (or many) more inclusive alternatives: “... *en el aula, el profesorado debería ...*” or “*en el aula, el cuerpo docente debería ...*” (no difference in English). A key factor throughout this research is how gender is used in the Spanish language. In this sense, Wasserman observes that languages where gender is grammaticalized, as is the case with Spanish, seem to imply the representation of two social classes, men and women, considering three assumptions: (a) “Reading languages with grammatical gender may prime people to express more sexist attitudes than when reading a language that does not have grammatical gender”,

(b) “Reading in a language with grammatical gender may make salient the historical oppression women have faced as a group”, and (c) “As a result, girls may rationalize the discrimination women have faced by expressing more sexist views” [1]. García Meseguer suggests two sources of linguistic sexism: Lexical, related to words, and syntactic, relating to constructs [2]. In this work, we focus on morphemes on gender.

The goal is to build an automated algorithm and to train it using an annotated dataset to learn when the use of a particular term might be considered as non-inclusive. After the algorithm is trained, it is used in the same context for validation, within a relevant amount of text, showing interesting results. To achieve this goal, a dictionary of terms that potentially might be used in a non-inclusive manner was generated, and relevant inclusive alternatives were defined for each term. A summary of the steps taken in this project is as follows:

1. A large dataset of documents is collected to be used for training and validation purposes. Selected documents are all electronic accessible doctoral theses created in Spain. Doctoral theses are usually very carefully written, deeply reviewed, and authors often have a reasonable culture background; that provides a better training dataset for our algorithm in terms of how difficult to find a non-inclusive usage is.
2. A dictionary of potential non-inclusive terms is generated:
 - a. Terms already present in several non-sexist writing guides (to the best of our knowledge, there is no dictionary of non-inclusive terms in the Spanish language).
 - b. Terms found in the document data set that are susceptible to being transformed to the feminine form; for instance, from “*profesores*” to “*profesoras*” (teacher). These were reviewed to rule out false positives.
3. The terms in the dictionary were located in the documents, and a representative context was extracted and stored (WIC, word in context, refer to section “Word in context identification” section).
4. All WICs were tagged using POS and transformed into an array of quaternions by mapping grammatical categories to integer numbers that algorithms can work with.
5. A non-linear SVM was trained with a split of the data set under supervision and tested against labelled data.

The paper is organized as follow: Section 2 provides the background found in literature and the gaps this research tries to fill; Section 3 describes the materials used in this project, especially the dataset, including its collection; Section 4 depicts the algorithm used, including alternatives evaluated; Section 5 shows how the annotation process was developed, in a four-step procedure until validation was performed; results are displayed in Section 6, evaluating the performance of the different alternatives tested, while Section 7 discusses results, and shows future lines of work.

2. State of the Art

The study of language in terms of gender inequality has attracted the attention of linguists over the past few decades. Major progress has recently been made in this area mainly for reasons related to equity, feminism, and even ideology. Several intergovernmental organizations and agencies (EIGE—European Institute for Gender Equality 2018—and United Nations 2020, among others) have newly developed toolkits and easy-to-use guides on how to use more gender-inclusive language [3]. However, some official institutions do not agree with the use of inclusive alternatives. In fact, Royal Spanish Academy (“Real Academia Española”, or RAE), the official institution and highest authority for the regulation of the Spanish language, established that the grammatical masculine in animate beings nouns is to be used to name all individuals of the species, without making distinctions between sexes [4]. Nevertheless, it is a controversial topic; in fact, several prestigious editorials, such as Elsevier, recommend the use of inclusive language for their publications [5].

The driving force in addressing the issue of language and gender research is the article *Language and Woman's Place* [6]. Since then, many scientists have continued to work in depth on this line of inquiry, demonstrating that discourse is a socially conditioned and institutionalized practice that reveals "meaning force and effect within a social context" [7]. Therefore, we may conceptualize this relationship in our case as follows: "if we take it that no expression has a meaning independent of its linguistic and non-linguistic context, we can plausibly explain the sexism of language by saying that all speech events in patriarchal cultures have as part of their context the power relations that hold between women and men . . ." [8]. Accordingly, language does not represent reality in a neutral manner but is rather a tool to strategically build the gender dimension in the public sphere [9–11].

There is ongoing work, therefore, that shows the link between women's social status and gender asymmetries in languages [1], [12] and stresses the need to analyze beliefs and discourse about men and women and how they are reflected in or compromised by language [7]. Having said this, we must not ignore that relationships between language and sexism are complex, as Cameron also acknowledges [8]. Many studies have been carried out on these aspects. For example, Newman et al. [13] focus on different uses of language by men and women; Foertsch and Gemsbacher [14] show how more and more frequently speakers and listeners, to combat prescriptivism, use the plural pronoun "they" to refer to singular antecedents to make language more inclusive. The requirements for the specification of referents' gender vary across languages and have further been explored in studies on several languages, including English [15,16], German [17,18], Swedish [19], Chinese [20–22], Polish [23], Italian [24,25], and French [26]. Other research has focused on the use of pronouns and their relation with gender marking [27–32].

In this vein, there have also been several research studies conducted with respect to the Spanish language [2], [33–35]. A great deal of current research focuses on whether Spanish is sexist or not. Cabello summarizes the arguments put forward by different authors [36]. On the one hand, those who argue that Spanish is not a sexist language put special focus on systemic and structural aspects [2,35,37], marginalizing the social dimension, while on the other hand, we have those who, placing emphasis on language as the creator of social reality, hold that Spanish is a clearly sexist language [38,39]. Almost all works and publications focus on the search for equity; that is, they try to find strategies to prevent discriminatory linguistic practices in terms of gender, race, etc. This search is associated with two strategies: Neutralization and feminization. In those languages in which the gender difference is not grammatically marked, the first strategy is used more often. The second is common when the objective is to provide more visibility to the feminine form. In the Spanish case, as reflected in practically all easy-to-use guides on how to use more gender-inclusive language, both strategies have already been used: Neutralization, for example, in those cases where generic nouns are used ("*persona*" instead of "*hombre/mujer*", or in English, person instead of man/woman); and feminization, where in the case of gender doublets, a generic feminine is proposed or the feminine forms of nouns that traditionally only presented masculine form ("*jueza*", "*médica*", etc.) are generalized (judge, medic).

Text mining refers to the process of extracting useful information from a document by identifying hidden patterns in unstructured text and has been a research subject for many different areas in recent years [40–43]. The analysis of the use of inclusive language in a text might be considered parallel to automatic sentiment classification, where opinion mining tries to extract subjective opinions from expressions. Most lines of investigation for text classification have been based on using a training set of samples to extract algorithms, including support vector machine (SVM) [44–46], k-nearest neighbor (kNN) [47,48], naïve Bayes (NB) [49–51], and decision tree [52,53], inside the classification paradigm [54–56].

Automated detection of specific language use has already been covered in previous studies, specifically on the problems of detection of hate speech [57], terrorism [58], racism [59], sexism [60], or any offensive language [61]. For the general problem of classification in text mining, [62] delves into classification techniques, and [63] combines several approaches to obtaining proper categorization. Support vector machines have historically been successfully used as a solution for binary classification

in text environments [64] under different approaches [65]. A common issue found through literature review is the problem of the bias that word embeddings suffer from, according to the text corpora the different solutions are built with. The subject has been studied and some research suggest that gender bias has not been solved yet [66]. Several attempts to remove that bias in the gender context stand out, including those that tackle the problem with a similar point of view as our research [67]; in [68], unintended bias is tackled from the misogyny detection perspective with reasonable results; in [69], authors claim to remove gender stereotypes while keeping reasonable and natural embeddings; in [70], bias in the hate speech detection context is quantified, and a novel knowledge-based generalization is proposed to remove that bias. Although most of the literature is focused on the English language, interesting approaches are also extending research to other languages such as French and Spanish [71,72]: They both show that word embeddings suffer from bias in hate speech and gender analysis, and different methods are proposed to overcome this issue. Despite the extensive research and to the best of our knowledge, no research has been carried out on the automatic detection of non-inclusive language in Spanish, whether it is in the academic production or any other context; the issue is different from hate speech or misogyny in that the exact same sentence can be considered as non-inclusive or not, according to a broader context. Finally, to the best of our investigations, no algorithm has been published to perform automatic detection of non-inclusive language in Spanish for these purposes.

In conclusion, the novelty of this work lies in the ability to detect, within a text, whether the words are used in a non-inclusive manner, based on the learning done in the training phase. Once trained, the algorithm can be fed with a document, and the non-inclusive expressions are found, based on their context. The research uses a wide dataset so that training effectively provides context to identify terms and compile them in a non-inclusive dictionary.

3. Materials

3.1. Technological Stack

The development was designed under Debian Linux using Pycharm (for Python text mining scripts, NLTK and Freeling), MS Windows 10 using versus Code (for C# validation software), and Debian virtual machines deployed in Azure for algorithm processing; hardware characteristics were based on a E16v3 Azure instance (16vCPU, 128GB RAM, 400 GB storage) for processing, and a DS2v2 modified Azure instance (2vCPU, 7GB RAM, 1,500 GB storage), for storing the documents and database.

3.2. Dataset

As for the scope, the initial goal is to use the automated decision maker inside the context of Spanish university academic and scientific production, where the texts to be analyzed are doctoral theses. In the 17 Spanish autonomous regions, there are, at the time of writing this paper, 73 universities, that have generated 257,564 doctoral theses; 102,914 of those theses were public domain and accessible, and after downloading and processing them, 100,450 theses were usable (the rest were not in Spanish language or in a legible format file). It is this area where the algorithm is to be designed, tested, and validated, with the thesis writers' age ranges and gender distribution presented in Figure 1.

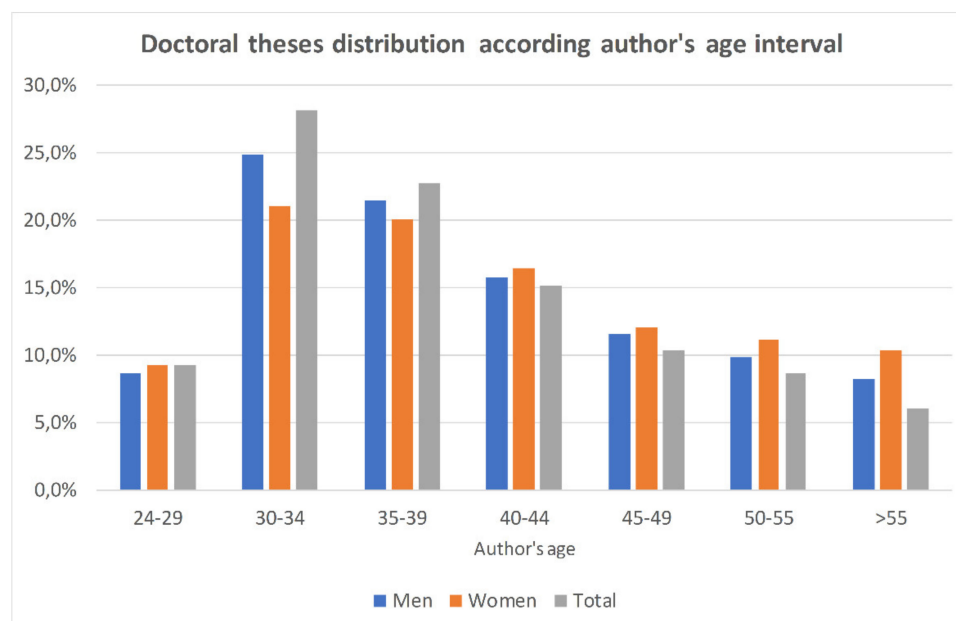


Figure 1. Distribution of women's and men's ages in doctoral theses production.

The distribution per autonomous region is shown in the following table, where Madrid and Cataluña are the most thriving regions of the country, and Andalucía the most populated, as displayed in Table 1.

Table 1. Doctoral theses per autonomous region.

Autonomous Region	Theses (%)
Madrid	25.9
Cataluña	17.4
Andalucía	15.8
Valencia	10.3
Galicia	5.2
Castilla y León	5.1
País Vasco	3.9
Aragón	3.0
Murcia	2.7
Canarias	2.4
Asturias	2.4
Navarra	1.9
Extremadura	1.2
Castilla La Mancha	0.9
Cantabria	0.9
Baleares	0.6
La Rioja	0.3

As for the subjects of the documents, they are organized under six categories (basics sciences, geosciences, biology and health sciences, engineering, social sciences, and humanities), that group the 24 epigraphs shown in Table 2.

Table 2. Doctoral theses per subject.

Subject	Theses (%)
Medical sciences	14.6
Life sciences	10.3
Technologic sciences	9.6
Chemistry	6.8
History	6.1
Economical sciences	5.5
Math	5.5
Physics	5.1
Psychology	5.0
Art	4.9
Sociology	3.4
Legal sciences	3.3
Pedagogy	3.2
Agricultural sciences	3.1
Linguistics	2.9
Earth and space sciences	2.8
Political sciences	2.5
Philosophy	1.2
Geography	0.7
Demography	0.6
Astronomy and astrophysics	0.5
Ethics	0.4
Logic	0.3

An additional requirement comes from the quantity of text. Table 3 summarizes the key data indicators compiled after initial investigations with regards to the context.

Table 3. Relevant indicators.

Characteristic	Values
Publication date	1974-2020
Average thesis/year	2232
Average page length/thesis	301
Average words/thesis	124,012

The doctoral theses created from April 1974 to February 2020 at 73 universities of Spain are the documents used in the dataset. They were treated in the same PDF format in which they are registered and range from 172 to 496 pages, with an average word count per document of 124,012. In this document corpus, we have found 12,457,005,400 words (9326 unique words), so care was taken at the design of the computing requirements and techniques to allow the handling of that amount of data (more than 12 billion elements). Moreover, since the aim of the project was to identify the non-inclusive use of term in modern Spanish texts regardless of the size of the text, special attention was paid to the methodologies used so that they would fit other use cases.

3.3. Document Collection and Storage

Doctoral theses from Spanish universities are stored in official repositories belonging to those public institutions. The first step consisted of not only downloading the documents and their metadata but also providing an automated system to obtain new documents as doctoral theses are uploaded without human intervention. University of La Rioja [73] maintains a repository that holds doctoral theses from Spanish universities (and others) and was used as source in this project. An ad hoc Python system service running inside a virtual machine was obtained for the official repository URL, showing the list of documents present, and compared with the documents already obtained (in the first run,

none). The differential (new) items were then downloaded, and their metadata were stored in a relational database present on a second virtual machine.

Along with the documents themselves, the scraper also obtained metadata that described the document and relevant information for later steps: Publication date, language, author, thesis title, contributors, and the UNESCO code that refers to the area of knowledge of the doctoral thesis; this information was saved in JSON format as show in Figure 2 and stored in a Mongo database.

```

"documents": [
  {
    "date": "2016-01-08",
    "language": "Español",
    "author": "XXXXXXXXXXXXXXXXXXXX",
    "title": "Desarrollo de metodologias para proyectos CubeSAT",
    "documentType": "doctoralThesis",
    "url": "http://hdl.handle.net/11093/684",
    "index": 42734,
    "UNESCO": {
      "3325.06",
      "3324.01",
      "1105"
    }
  }
]

```

Figure 2. Document object representation.

Plain text was obtained from the documents using several libraries in the Python language related to text extraction and treatment and a natural language processor. The words extracted were again stored in a database, properly configured to be able to store a large amount of information. Once stored, extracted words were transformed according to a set of rules that converted them into a numerical representation, that allowed the generation of the classification algorithm. A lexical analysis was completed to tokenize and separate the lexical components of the text (isolating lexical separators such as white spaces and punctuation signs). The text was also curated by eliminating words that, given their location in the document, would not receive the non-inclusive use check: Tables of contents, text inside tables, text inside pictures and diagrams, formulae, numbers, units, words in languages other than Spanish (checked on a Spanish word dictionary), references, and page numbers. Although section titles might display a non-inclusive use, their presence in the table of contents was skipped to avoid text repetition. A morphological tag was also assigned to every lexical component in a process known as part-of-speech (POS) tagging, in which every lexical component receives a characterization consisting of a type (article, substantive, verb, adjective, adverb, etc.), gender (masculine or feminine), number (singular or plural), etc. Finally, the text was again stored in the database related to the doctoral thesis it came from. Duplicate words were not filtered out, since many repetitive ones like definite-indefinite articles take a key part in the WIC as well as in the non-inclusive detection.

4. Algorithm Design

Classification is a well-known problem: A sample is obtained; the unstructured data (text) are organized in a structured format; every example is measured in the same way, and the answer is expressed in terms of true or false, a binary decision. In mathematical terms, a solution is a function that maps examples to labels, $f: w \rightarrow L$, where w is a vector of attributes and L is a label (for supervised learning) [74]. In this case, the attributes are words' grammatical characteristics, and the label is the inclusive or non-inclusive use of the term in the context. Given there are two alternatives (inclusive and non-inclusive) for the classification, we are at the boundaries of binary classification; Figure 3 displays the process, where dictionary terms susceptible to being used in a non-inclusive way were populated using the reviewed part of the document dataset; the dataset was transformed into numerical vectors to feed into the algorithm and be used as classifiers.

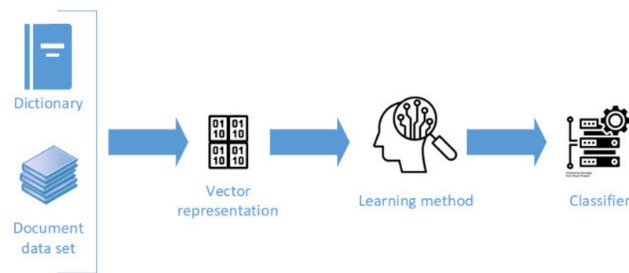


Figure 3. Proposed workflow.

In supervised classification, the model is tuned to the training samples, which are a portion of the dataset to be observed. The algorithm is based on the vectors that are already labelled and generate a predictor for future (unlabeled) cases; although such algorithms may suffer from issues (e.g., overtraining and dependency on the similarity of the labelled samples and the dataset), they have been successfully used in many different scenarios [75]. Figure 4 depicts classification process.

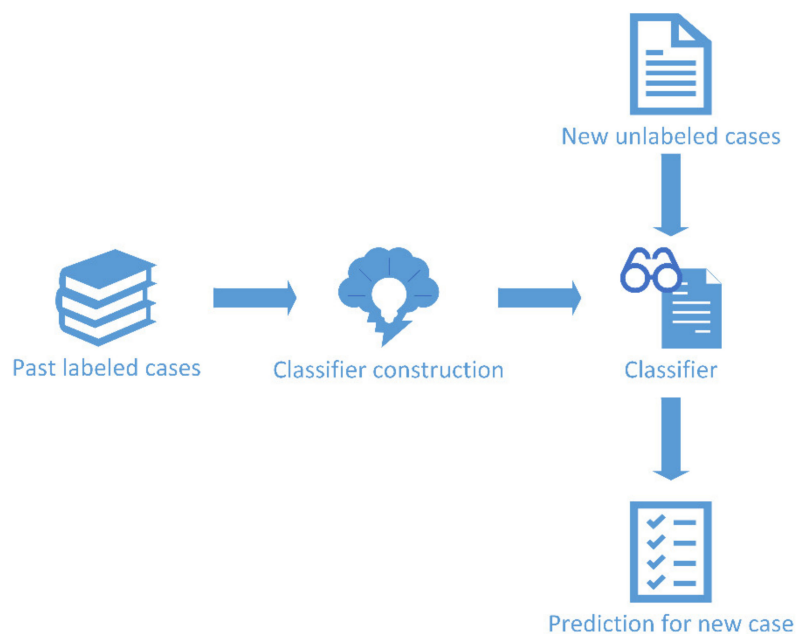


Figure 4. Classification process.

In the case of text mining, the words in the document are the characteristic features; therefore, the text must be transformed into numerical values the algorithms can work on. Moreover, there are terms that are susceptible to being used in a non-inclusive manner and others that will not be used in any context. Finally, even terms that are considered non-inclusive in one context may not be considered so in a different context.

With our requirements in mind, the goal was to find a machine learning classification algorithm that would identify (classify) whether the use in its context of every term in the dictionary we compiled is inclusive. After reinforcement learning and unsupervised learning were considered, supervised learning was the chosen option. First, the algorithm was given insights on a small training set of documents (for learning when a term is inclusive or not in context) and then it was fed with the test set of documents (for tagging terms in different contexts and validating the model). After a group of linguistic experts labelled the text words of a subset of the documents as inclusive or not based on the relative contexts, a 10-fold cross-validation method was implemented to avoid under-fitting or over-fitting. To prevent the results of the learning phase analysis from being independent of the subset chosen, document collection was divided into two parts: A training set for the machine learner to build

the model and a test set to be validated. Every word that might plausibly be used in an inclusive or non-inclusive way (whether found in the learning step, or susceptible to being used in both genders) was then analyzed by the algorithm to predict a discrete binary property (non-inclusive or inclusive). Figure 5 shows an overview of the training process.

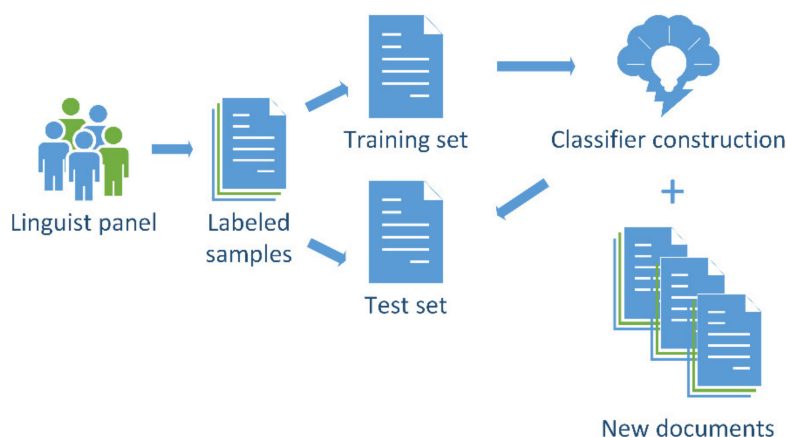


Figure 5. General overview of the classification training using two subsets of labeled samples for generating the classifier.

For building the classification model, the classification method used was a support vector machine. Previous authors' projects in the area of predicting intentions have shown SVM (with a reasonable false-positive and true-negative ratios) was a successful mechanism to build classification algorithms; moreover, those experiences provided certain methods and library functions that made SMV something authors were confident and comfortable with. SVM, considered an extension of the perceptron algorithm, has the optimization goal of setting a decision line that marks a frontier between the classes of data and maximizing the margin between that frontier and the sample points closer to this hyperplane; those points are, in fact, the support vectors. Each data entry (belonging to the set of samples for learning) is represented as an n -dimensional vector, where every component is a number.

The mission of SVM is to find the hyperplane that yields the largest minimum distance to the points of the training set (the margin), using a small subset of vectors from the training set. Whether in a linear or non-linear problem, SVM separates the data into two classes (in our case, non-inclusive or inclusive) by mapping the information into spaces with more than two dimensions. Once the model was created, the test data were incorporated, and the classification supervised, as shown in Figure 6:

Supervised classification is widely used within either pure statistics (logistic regression, discriminant analysis) or artificial intelligence paradigms (neural networks, decision trees, Bayesian networks) [76,77]. At this stage, a set T of n training feature vectors $X_i \in R^D$ was separated, where $I = 1, \dots, n$, and the corresponding class labels $y_i \in \{+1, -1\}$ (for the binary non-inclusive/inclusive classification). Sample vectors labelled as inclusive have the class label $+1$ (are class $C+$, positive values), and the rest are non-inclusive (belong to the negative class $C-$).

The linearity of the problem (existence of a hyperplane function defined by $x \cdot w + b = 0$ that keeps the maximum distance between the inclusive or non-inclusive classes) was discarded after initial convergence tests. The characteristics of the WICs in our cases do not allow us to find the w vector that would maximize the distance from the binary classes in the form of

$$w = \sum_{i=1}^n \alpha_i y_i x_i \quad (1)$$

with the α_i parameters solved by the function

$$\max \alpha \left[\sum_i \alpha_i + \sum_{i,j} \alpha_i \alpha_j y_i x_i \cdot x_j \right] \tag{2}$$

where the α_i variables are the Lagrange multipliers and C is a parameter that penalizes WICs that do not have a correct classification; after iterating through different values of C (beginning in C = 1), C = 145 was selected for showing the lowest cross-validation error:

$$\sum_i \alpha_i y_i = 0 \quad 0 \leq \alpha_i \leq C \tag{3}$$

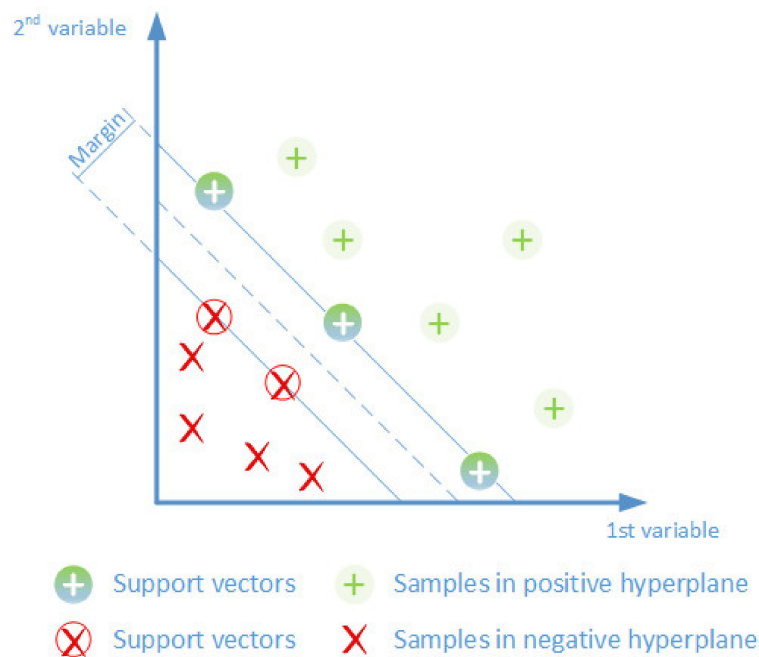


Figure 6. A two-dimensional representation (for clarity) of how support vectors help classify new samples by vicinity in the hyperplane.

When dividing the two groups with an n-dimensional hyperplane is not possible, since the data points are separated by a nonlinear region, classification cannot be obtained by simply calculating the direct inner product between the points. The non-linearity can be solved by using a kernel function to map the information to a different space and then perform separation (kernel functions construct non-linear decision surfaces for sample classification). According to the Hilbert–Schmidt theorem, a symmetric operation that meets Mercer’s condition can be represented as if it were an inner product when

$$\int \int K(x, x') \varphi(x) \varphi(x') dx dx' > 0 \tag{4}$$

when $\varphi(x) \neq 0$ and $\int \varphi^2(x) dx < \infty$.

In this way, the problem is rewritten to maximize

$$\sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{5}$$

where $\sum_{i=1}^n \alpha_i y_i = 0$ and $0 \leq \alpha_i \leq C$, resulting in an optimal classification function as:

$$f(x) = \operatorname{sgn} \left[\sum_{i=1}^n \alpha_i y_i K(x_i, x) + b \right] \quad (6)$$

Two non-linear kernels were tested in the preliminary phases with 109 WICs to test suitability (RBF and polynomial). The sigmoid was discarded because of convergence and computation issues. Thus, the functions used were as follows:

An RBF (radial basis function) kernel, which has the property that each basis function depends only on the radial distance from a center, written as:

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2) \quad (7)$$

where $\gamma > 0$.

A poly(nomial) kernel, which is directional (the output depends on the two vectors in the low-dimensional space because of the dot product), following:

$$K(x_i, x_j) = (\gamma x_i^T x_j + r)^d \quad (8)$$

where $\gamma > 0$.

5. Annotation Process

In this section, the steps performed on the dataset until the whole system was validated are described, as shown in Figure 7.

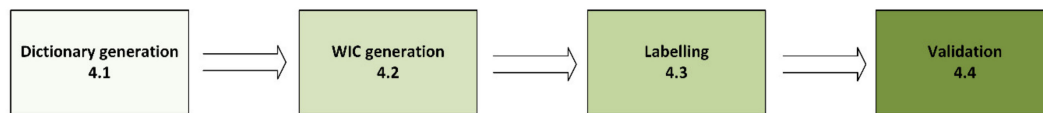


Figure 7. Summary of the process.

5.1. Generation of the Dictionary of Potential Non-Inclusive Terms

In this section, the process for the generation of the dictionary of terms with potential non-inclusive use is depicted. Two sources for that dictionary were used: First, a list of terms that, found in the dataset, matched a rule to be a candidate; second, a list of terms compiled from different non-inclusiveness guides. The process is shown in Figure 8.

Considering the quantity of words involved (more than 12 billion), the size of the training subset exceeded the human resources available for tagging the data. For this reason, a dictionary of terms that potentially might be used in a non-inclusive manner was created to guide and speed up the process. The words for the dictionary came from two different sources: First, a compilation of different guides on inclusive use of language [78–80] and second, terms that were found in the dataset used in their masculine form and that were located in their feminine form as well. This second group of terms represented those words that match one of the generic rules for inclusive language: First, the case of “generic masculine” terms, such as “*usuarios*” (users), which should come with their feminine form (“*usuarios y usuarias*”, no difference in English) or be replaced with a truly generic alternative “*personas*”, (persons); second, the case of abstract substantive terms, such as “*alumnado*” (student body) instead of “*alumnos*” (student); third, the case of metonymic substantive terms, which refer to the position/profession/activity of a person: “... *la edad para el ciclista puede ...*” (the age of the cyclist) versus “... *la edad para practicar el ciclismo ...*” (the age to practise cycling); and fourth, the case of the adjectives in the past participle, where a gender neutral noun is inserted: “*los ciegos*” versus “*las*

personas ciegas” (blind people in English for both cases) or the wording changed: “ ... *todos los que quieran* ... ” versus “ ... *quien quiera* ... ”, (who wants in English for both cases).

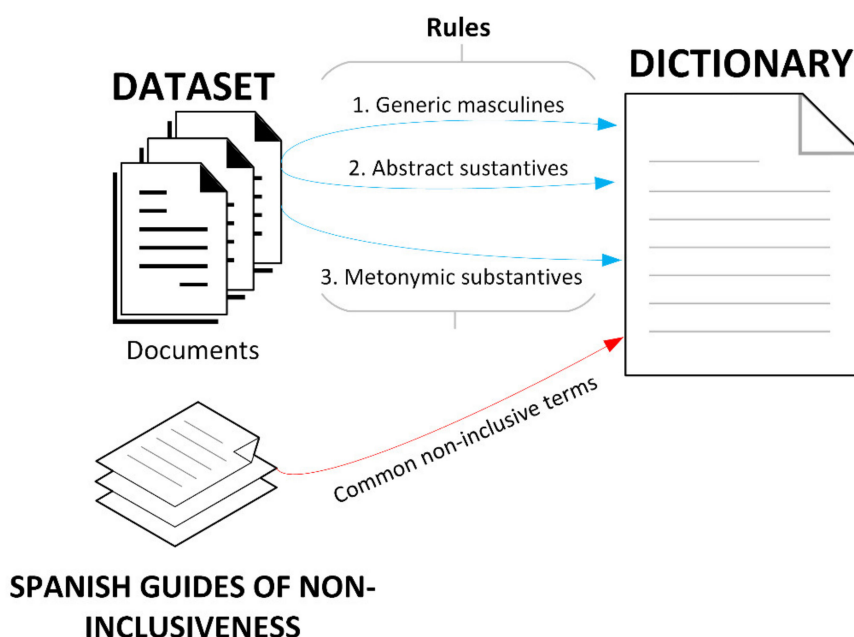


Figure 8. Generation of the dictionary of potential non-inclusive terms.

To find the second group of terms to be added to the proposed dictionary, the following process was performed. First, every noun: “*niño*” (boy), “*profesor*” (teacher), and adjective: “*correcto*” (right), “*incorrecto*” (wrong) in masculine form was obtained from the documents, and then alternatives were selected to generate a list of unique terms. Then, stemming and lemmatization was performed to normalize words to a single form; this is to choose the word that has a match in the Spanish dictionary; for instance, from plural “*profesores*” (teachers) to singular “*profesor*” (teacher). Once normalized, a feminine form was looked for: Considering the rules of feminine construction in Spanish, the Freeling library with the Spanish tag set was checked to determine whether any combinations would create a valid term as shown in Table 4.

Table 4. Generating feminine.

Case	Rule
Trailing “e” or “o”	Substitute with “a”
Trailing consonant	Add “a”
Trailing “ista”	No change
Trailing “dor”/”tor”	Change to “triz”
Any	Add “esa”/”isa”/”na”/”ina”

Once the dictionary was created and the unique terms isolated, it was reviewed by a panel of experts in Spanish linguistics, who ruled out 13.8% as false positives: Words that matched at least of the feminine generation patterns but are not susceptible to non-inclusive use. It is important to mention that new forms of inclusive language including “wildcard” characters to abstract the word from being masculine or feminine (as “*todes*”, “*tod@s*”, “*tod*s*” or “*todxs*”) were explicitly looked for, but not found in any document as part of the dissertation.

5.2. Word in Context Identification

In this section, the transformation of the dataset into a list of potentially non-inclusive terms with their broad context is addressed. As displayed in Figure 9, every word in the dictionary was looked for

in the documents in the dataset, and if found, it was extracted with a relevant context that will allow later identify the inclusiveness or not of its use.

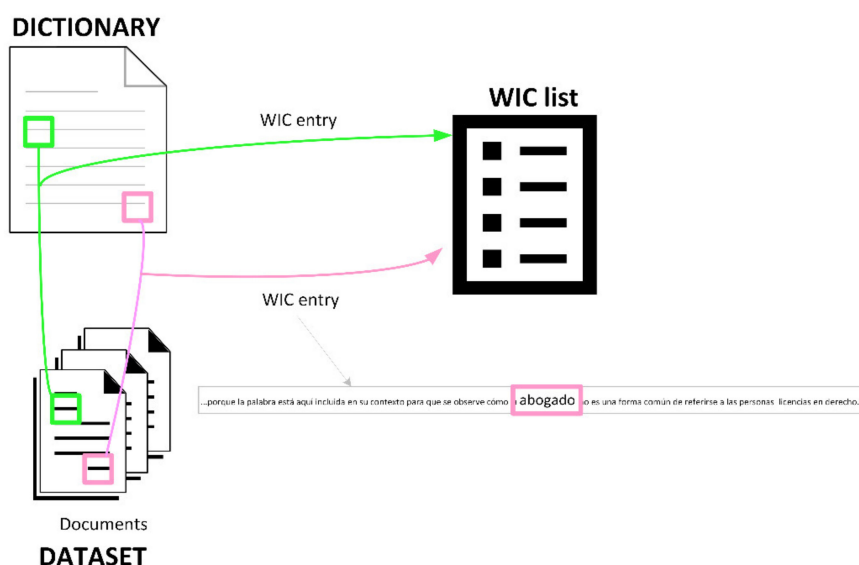


Figure 9. Generation of the word in context list.

Since the characteristic of being non-inclusive applies not to the existence of a term but to its use in a specific context, for every word included in the dictionary, every document in the dataset was scanned for that word. When found, a data structure was generated using the following rules:

- The 17 words preceding and 17 words following the located dictionary term were initially taken to generate the word in context (WIC), having thus enough context to keep its meaning. “... a bolsa de 2002. El equipo de Alemania pasa de estar constituido como asociación propiedad de sus socios a ser una sociedad mercantil cotizada en bolsa con la propiedad totalmente diluida. Este ejemplo no fue ...” (‘socios’ is the term in the dictionary, and the rest of the words are the 17+17 context); in English, (... to 2002 stock exchange. The team in Germany went from being incorporated as an association owned by its partners to being a publicly traded trading company with fully diluted ownership. This example was not ...).
- The WIC was adjusted to a sentence, so that when a period was found, the trailing words were discarded. “El equipo de Alemania pasa de estar constituido como asociación propiedad de sus socios a ser una sociedad mercantil cotizada en bolsa con la propiedad totalmente diluida.”
- The WIC was characterized into lexical categories by classifying and labelling the words in it with the NLTK Stanford POS combined taggers for the Spanish language:
 - Unigram and bigram taggers were trained.
 - The bigram tagger tried to tag every token in the WIC. If unsuccessful, unigram or default taggers were tried sequentially.
 - The results were evaluated to check for tagging success, and in that case, the WIC was stored as the sequence of ordered tagged elements (using the EAGLE tag-set): [(‘EI’, ‘da0ms0’), (‘equipo’, ‘ncms000’), (‘de’, ‘sps00’), (‘Alemania’, ‘np00001’), (‘pasa’, ‘vmip3s0’), (‘de’, ‘sps00’), (‘estar’, ‘vmn0000’), (‘constituido’, ‘vmp00sm’), (‘como’, ‘cs’), (‘propiedad’, ‘ncfs000’), (‘de’, ‘sps00’), (‘sus’, ‘dp3cp0’), (‘socios’, ‘ncmp000’), (‘a’, ‘sps00’), (‘ser’, ‘vsn0000’), (‘una’, ‘di0fs0’), (‘sociedad’, ‘ncfs000’), (‘mercantil’, None), (‘cotizada’, None), (‘en’, ‘sps00’), (‘bolsa’, ‘ncfs000’), (‘con’, ‘sps00’), (‘la’, ‘da0fs0’), (‘propiedad’, ‘ncfs000’), (‘totalmente’, ‘rg’), (‘diluida’, ‘vpm00sf’)]

- The WIC was translated to a vector of false quaternions and stored, to generate, in the next step, the sample vector the algorithm uses to predict the non-inclusive/inclusive classification of the WIC. Every word is given four properties: Category, type, gender, and number. Each property is assigned an integer value according to Tables 5–7, depicting the mapping.

Table 5. Mapping rules for categories and types.

Category	Value	Type	Value
Adjective	1	Qualificative	1
		Ordinal	2
Adverb	2	General	3
		Negative	4
		Determinant	3
Determinant	3	Demonstrative	5
		Possessive	6
		Interrogative	7
		Exclamative	8
		Undefined	9
		Article	10
		Noun	4
Verb	5	Proper	12
		First-person	13
Pronoun	6	Second person	14
		Third person	15
		Personal	16
		Demonstrative	17
		Possessive	18
		Undefined	19
		Interrogative	20
Relative	21		
Exclamative	22		

Table 6. Mapping rules for gender.

Category	Value
Feminine	1
Masculine	2
Common	3
Neutral	4

Table 7. Mapping rules for number.

Category	Value
Singular	1
Plural	2
Invariable	3

Thus, for instance, the word “*equipo*” (team) would map to the vector [4,11,2,1], being a common noun that is masculine and singular. This rule was applied to every word in the stored WICs.

After the validation of 617 WICs with an average sentence length of 26 words, it was determined that the necessary context was given by extending the WIC 17 words both left and right, centered on the non-inclusive term. The algorithm was also tested in different contexts having less average length (general audience magazines, newspapers, blogs) and others with greater average length (legal documents).

The summary of the more relevant information stored as a result of this step is as follows. The following list illustrates data structure for storing information:

- Non-inclusive term: The word that is used in a potentially non-inclusive way in the found context.
- Where used: The group of contexts where the non-inclusive term has been found in the dataset:
 - Document ID: The document where it was found.
 - WIC: The different contexts in the present document.
- Starting point: The word number inside the document where the WIC begins.
- Length: The number of words of the WIC.
- Text: The text of the WIC.
- POS: The EAGLE tagging for the WIC.
- Inclusive: Whether the term is considered to be used in an inclusive manner or not; this is informed with human intervention in the training (next) step or marked by the algorithm.
- WIC correction: In case the use of the term is marked as non-inclusive, an alternative writing using an inclusive replacement is given, including (potentially) a modification of the WIC (substituting some terms with alternatives).
- Inclusive alternatives: A list of alternative inclusive terms with example WICs, obtained in the training step.
- Sample vector: The vector of properties for the WIC, containing the information used for labelling (training) and prediction.

5.3. Labeling

In this section, the process of labelling the elements in the WIC list is described, as well as how that labelling helps building the classification algorithm. On the WIC list, two actions are performed on it. First, it is transformed into a mathematical representation according a series of mapping rules. Second, it is labelled by a panel of three teams that annotate WIC elements divided into three groups; part of the elements are distributed to two different teams, and whenever a disagreement in the evaluation was found, it was also evaluated by the remaining team, to decide final evaluation. The mathematical representation of the WIC and the labels, all together, lead to the labelled matrix of features and the binary characteristic, foundation for the SVM to build the classification algorithm. The whole process is shown in Figure 10.

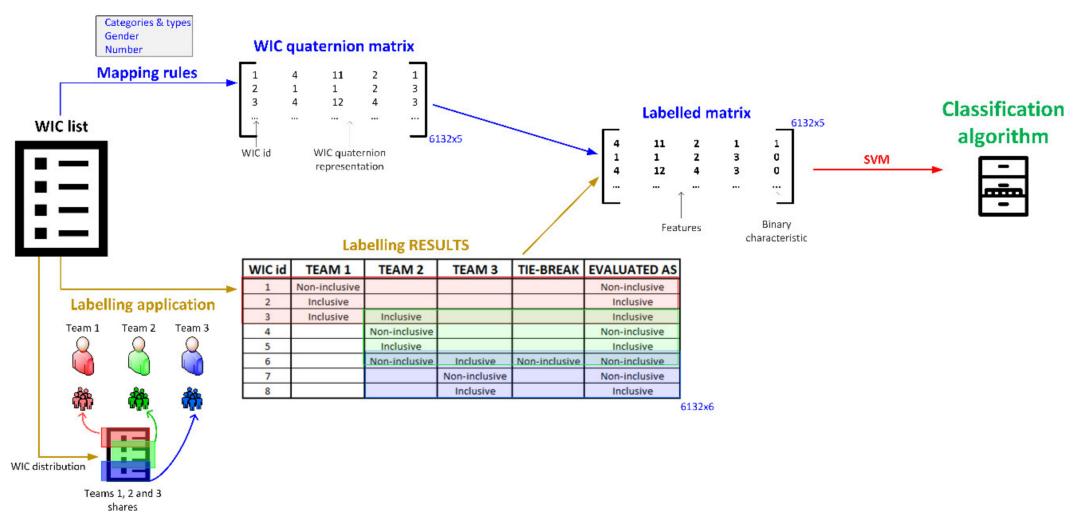


Figure 10. Sample labelling, that leads to the generation of the classification algorithm.

Once the WICs for each entry in the non-inclusive terms dictionary had been identified and tagged, 6132 WICs out of the documents in the dataset were labelled for algorithm training and validation purposes. The WICs corresponded to 20 WICs per term in the non-inclusive dictionary. For the

ultimate goal of projecting the results of the learning step to the unlabeled samples, 20% of the labelled samples were retained for validation and error measurement.

The WICs were presented for labelling in an ad hoc application as shown in Figure 11 that allowed several linguistics to simultaneously classify each of them into one of the binary categories: Inclusive use or non-inclusive use. For that purpose, our data structure was modified with the “inclusive” Boolean value and the sample vector (X_i, y_i) , where X_i is the p -dimensional feature and y_i represents the binary inclusive characteristic that is evaluated. Whenever a term in the non-inclusive dictionary was found in a document, its relevant context was extracted and shown, providing a means for linguists to validate its inclusivity or to supply a re-written alternative and select an inclusive term.

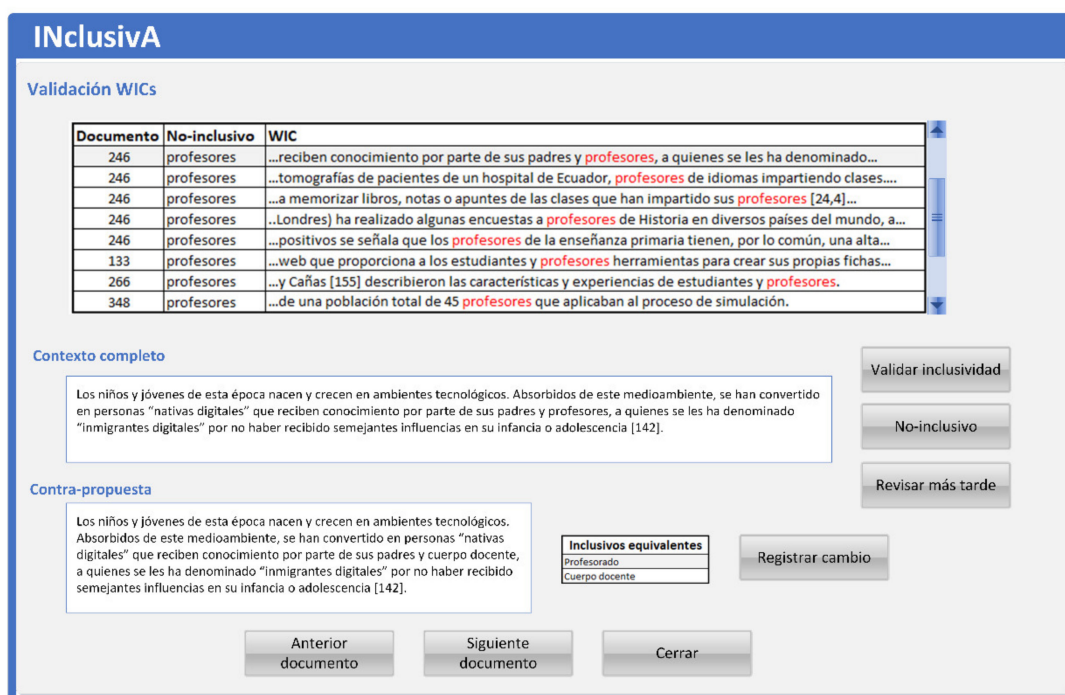


Figure 11. The ad-hoc C# application used for word in context (WIC) labeling in the training phase.

Out of the labelled WICs, 4906 were separated for training purposes (corresponding to the 80% not reserved for validation). Each of them was categorized into one of the two classes (inclusive or non-inclusive). The process was performed by three teams of linguists, with their assistants; the WICs were distributed randomly among the teams: 1535, 1535, and 1536 unique WICs (labelled by one group) plus 300 shared WICs, that were split in two groups to be analyzed by two teams at a time. Table 8 summarizes process information.

Table 8. Annotation process.

Group	Group 1	Group 2	Group 3
Unique WICs	1535	1536	1535
Shared WICs	150	150	-
	-	150	150
Total WICs per group	1685	1836	1685
Members	2	4	2

Labelling was carried out in three steps: In the first step, the members of all groups agreed on the non-inclusive language rules to be used, which resulted in the considerations in Section 5.1; in the second step, shared WICs were labelled using a shared spreadsheet containing WIC number, WIC’s text, and the classification given by two different groups (“I” for inclusive, “NI” for non-inclusive); in

the third step, every WIC that had been classified differently in step two was labelled by the group that had not yet given label for that WIC. The final classification for the WIC was given, consequently, by two votes (when there was initial agreement), or three votes (should there had been a different classification). Of the 300 shared WICs, nine discrepancies were found: Seven needed the third vote to achieve final classification, and two were classification mistakes by human error (that turned out to be agreed after revision); of the 4906 classified WICs, 3129 were finally labelled as non-inclusive, which represents 64%. Figure 12 displays selected records to illustrate the different cases. Columns show WIC number, and classification given by groups one, two, and three (“I” for inclusive, “NI” for non-inclusive). The fourth column shows final classification, that is “I”/“NI”. That final classification was straightforward for cases where the two votes were aligned (record 121) or needed an extra vote (record 162).

WIC	G1	G2	G3	RESULTADO	Texto
121	I	-	I		...federativas en España (RFEF, 108). A nivel mundial, estima que hay más de 12 millones de jugadores y jugadores de fútbol sala. Sin embargo, este crecimiento natural...
162	I	NI	NI	NI	Reposición de líquidos y su efecto sobre niveles de deshidratación en jugadores de fútbol sala en función de la posición ocupada en el terreno de juego. Se puede ver ...
163	NI	NI	NI		Suplementación antioxidante de 6 semanas como estrategia frente al daño oxidativo en jóvenes jugadores de fútbol sala. Aparentemente no tiene efectos secundarios ...

Figure 12. Spreadsheet used to manage classification.

5.4. Validation

In this section, validation of the whole process is addressed. The WIC list is split into two groups: 80% for training the algorithm and 20% for validation. That 20% group of labelled samples is applied the prediction algorithm to classify the use of the term as inclusive, or not; that computerized predicted value is then compared with the label provided by annotators in the previous step (human-made labelling), so that error, accuracy, and other key indicators can be obtained to evaluate the prediction. Figure 13 depicts this process.

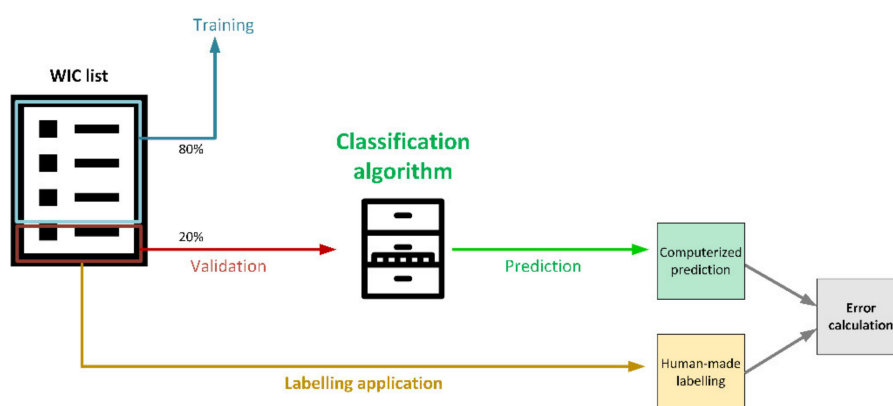


Figure 13. Validation process.

As in many classification problems, one of the most important performance measures is the classification accuracy (ratio of terms classified the right way versus detected cases). Nevertheless, extra indicators were added since there is a certain asymmetry in cases of classification failure: Detection of non-inclusive use of language where this issue is not happening has a greater impact than does the failure to detect non-inclusive terms because in the event of the former, we may lose users’ confidence from the misclassification. For this reason, the type of classification error was also added as an additional key indicator to evaluate performance. To obtain both parameters, the subsequent notation was used, as shown in Table 9.

Table 9. Calculating performance.

Indicator	Variable
Number of inclusive WICs to be classified	n
Number of non-inclusive WICs to be classified	m
True positives	II
True negatives	NN

It is assumed that a false positive is obtained when a WIC is classified as inclusive but it is actually non-inclusive; a false negative occurs when a WIC is categorized as non-inclusive but is actually inclusive; and true positives and true negatives are inclusive and non-inclusive detections that are not misjudged. Table 10 defines the notation used.

Table 10. Notation definition.

Term	Abbreviation	Detected by Algorithm As	Labelled by Linguistic As	VariableName
True positive	TP	Inclusive	Inclusive	II
True negative	TN	Non-inclusive	Non-inclusive	NN
False positive	FP	Inclusive	Non-inclusive	IN
False negative	FN	Non-inclusive	Inclusive	NI

According to this notation, selected key indicators can be written as follows:

$$\text{Classification error (E)} = 1 - A = \frac{IN + NI}{n + m}$$

$$\text{Classification accuracy (A)} = \frac{II + NN}{n + m}$$

Based on these parameters, two other key indicators were defined: FPR and FNR. The false positive rate (FPR) is the ratio between the WICs categorized as inclusive when there is a non-inclusive usage (number of false positives, or FP) and the ground truth negatives (where ground truth negatives are the sum of true negatives, TN, plus false positives, FP); the false negative rate (FNR) is the ratio between the WICs categorized as non-inclusive where there is in fact a normal inclusive usage (number of false negatives or FN) and the ground truth positives (where ground truth positives are the sum of true positives, TP, plus false negatives, FN).

$$FPR = \frac{FP}{FP + TN} \quad (9)$$

$$FNR = \frac{FN}{TP + FN} \quad (10)$$

Finally, three other indicators were included to evaluate classification: Precision (P, confirmed positive class predictions), recall (R, to show missed positive predictions), and F-measure (balance between precision and recall in one indicator); according to expected class imbalance, micro-average was selected to compute the indicators:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (11)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (12)$$

$$F\text{-measure} = \frac{2 * P * R}{P + R} \quad (13)$$

6. Results

After the validation step, where the dataset of unused labelled sample points reserved for verification were utilized for testing performance, the values displayed in Table 11 for the key indicators were calculated following the formulae shown in the previous section.

Table 11. Evaluated performance.

Indicator	RBF	Polynomial	Baseline
	Value	Value	Value
Error	11.1825%	14.3175%	25.5714%
Accuracy	88.8175%	85.6825%	74.4216%
False positive ratio	10.1299%	11.1%	23.0404%
False negative ratio	12.2137%	13.9286%	28.5714%
False positives	39	58	97
False negatives	48	62	102
Validation labelled sample points		778	
Inclusive		357	
Non-inclusive		421	
Precision	88.7931%	83.5694%	72.4432%
Recall	86.5546%	82.6331%	71.4286%
F-measure	87.6596%	83.0986%	71.9323%

RBF seemed to perform better than polynomial, especially on false negatives (the most concerning point). Using RBF, consequently, the accuracy indicator of the algorithm reached nearly 90%, fulfilling initial expectations (considering that a certain quantity of classification errors was inevitable). Despite the fact that null classification was not achieved, it is the false negatives (terms used properly in an inclusive way that were wrongly detected as non-inclusive usage, making up 6.1696% of the manually labelled samples reserved for validation) that will require greater attention for this algorithm to work in a production environment. False positives are more easily addressed, as a 0% error rate is not to be expected, and undetected non-inclusive usages do not cause loss of confidence: False detections actually occur only when users experience such losses.

The third column displays a comparison of the SVM classifier against a baseline that was also performed. Having certain imbalanced distributions between our binary classes, majority class was selected to label every test instance to the majority class in the test set. As expected, the algorithm performed better than the baseline.

In terms of the most common non-inclusive terms found in the whole dataset, Table 12 summarizes the top 80% of the phrases found in the WICs categorized as non-inclusive, once a potentially non-inclusive term was detected:

Table 12. Most common non-inclusive terms.

Non-Inclusive	Frequency		Inclusive Equivalent
	Occurrences	%	
"Alumnos" (pupils)	273,224	16.09	"Alumnos y alumnas"
"Hombre" (man)	259,300	15.27	"Ser humano"
"Estudiantes" (students)	241,470	14.22	"Estudiantado"
"Profesores" (teachers)	186,112	10.96	"Profesorado"
"Jóvenes" (youth)	169,980	10.01	"Juventud"
"Directores" (directors)	137,886	8.12	"Dirección"
"Niños" (children)	52,301	3.08	"Niños y niñas"
"Profesionales" (professionals)	23,604	1.39	"Personas profesionales"
"Clientes" (customers)	16,641	0.98	"Clientela"

Statistics on the doctoral theses by year of publication showed an unexpected evolution in the last two years of our sample period. In Figure 14, the number of documents by year is shown (blue line), and the orange line shows the number of non-inclusive terms found in those documents per year at an adjusted scale. For the rest of the period under study, the percentage of non-inclusive terms generally followed the overall trend in the number of documents (that is, more documents imply more cases of non-inclusive usage of terms).

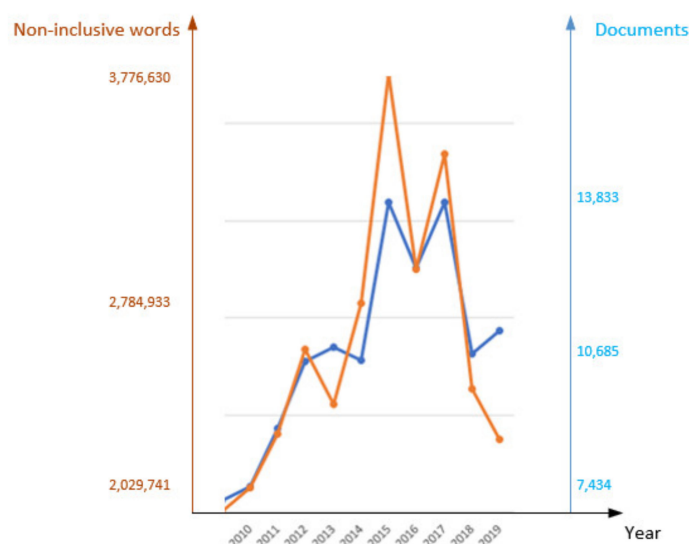


Figure 14. Comparative evolution of documents and the detected non inclusive uses of terms.

That general rule applies until the last two years under study (2018 and 2019), where the ratio of non-inclusive usages found per document declined by 36.3% versus the average of previous years. Perhaps the explanation can be found in the current gender equality politics and regulations put in place by the Spanish universities to raise awareness of gender-sensitive language.

After reviewing the corpus of doctoral theses, we verify that the sexist traits in the analyzed language area are less frequent. Therefore, despite aspects such as the use of the generic masculine or the low social acceptance of the feminine grammatical gender of some nouns, the results show that awareness is increasing, something fundamental considering the influence that can be exerted by academia on society.

7. Discussion and Future Work

This research represents the first step taken to provide automated tools dedicated to the use of inclusive language for the Spanish language: the goal is to identify, within a text, if the words are used in a non-inclusive manner (on the basis that the words are all inclusive, but it is the way of using them that might be non-inclusive). The approach involves the use of a dictionary of potential non-inclusive words, the categorization of the “suspicious” words found in their context, and an SVM classifier to categorize whether the use of that word in that context can be considered as non-inclusive. The dataset used to train, validate, and test the system was the digital accessible Spanish doctoral thesis (more than 100,000 documents, averaging 300 pages each file), considered wide enough for this purpose.

The results of this research are in two directions: First, the algorithm itself, that can be applied on other type of Spanish documents to detect non-inclusiveness with a small false-negative ratio; second, the analysis of the most common words used in a non-inclusive way, to be fed to many governmental inclusive awareness actions. Two further projects are expected to be executed as continuation: First, the creation of a Python library with the trained algorithm to perform this type of classification on any Spanish text; second, the publication of the dictionary of more frequent terms used in a non-inclusive

manner, with its inclusive alternative recommendations, based on the preliminary one built within this project.

To evaluate the dictionary, a final test was performed. Ten doctoral theses (2836 pages, for a total of 1,187,712 words) were selected randomly, but using the following criteria: First, written in the last eight years; second, half of them written by men, the other half by women; third, their topics distributed according to the percentage of subjects in the entire dataset (Table 2). The documents were labelled by the algorithm, and then manually tagged to discover non-inclusive usages of terms that were not in the dictionary (and consequently, that had not been found by the algorithm). In that experiment, three different terms were found (as shown in Table 13).

Table 13. Evaluating dictionary.

Term	Term (English)	Non-Inclusive Usages
Cineasta	Film maker	7
Corresponsal	Correspondent/journalist	9
Guía	Guide	13

The WICs in which those terms were found are 422, of which 29 were considered as non-inclusive, and the rest were labelled as inclusive. The case of those words showed that dictionary has a solid basis (three in more than 1 million), since they are quite exceptional (they use the same form for masculine and feminine and are exceptions to the standard rules described in Table 4).

Results showed that algorithm would also detect non-inclusive uses of the language that are not related to masculine forms of nouns and adjectives. For instance, the case of “... el estudiante de una lengua extranjera ...” (the student of a foreign language); the word “estudiante” (student) can be used for the masculine and feminine form. In this case, the algorithm detected this use as non-inclusive because of the determinant in the noun phrase, not the kernel of the noun phrase. The main limitation of this approach comes from the fact that it needs to be given a suitable context for a word to be detected as used in a non-inclusive manner. In order for it to be used to validate inclusiveness in very short sentences given with no further context, it will not find enough background to detect non-inclusive used: For sentences like “los profesores van a venir” (“the teachers are going to come”), with no further context, no classification is obtained. Apart from that, there are still many relevant fields to be further investigated and improved:

- Other well-known solutions for the kernel non-linear approach to the problem may help to reduce the error rate and specially eliminate the false negatives issue. Different alternatives shall be tested and compared to find a solution that optimizes convergence, computing time, and a minimum false negative rate.
- In this research, every doctoral thesis was considered to be an independent sample point regardless of its origin. It is expected that according to certain characteristics related to the origin of the document (gender and age of the author, knowledge area of the document, and/or the internal department to which the author is attached, date of publication, etc.), an asymmetry in the results would arise. Another study will try to find correlations between a document’s external properties and the ratio of non-inclusive terms per total number of words in a doctoral thesis.
- The document set used in this project was made of doctoral theses. Another study should be put in place to compare results in a non-academic environment, and check if the most common non-inclusive terms match the ones got int this research.
- Although for creating and training the algorithm performance was not an issue, we consider that computation time should be reduced so that actual rate analysis (16 days to analyze the whole dataset) could be improved and adapt its performance to an online “on-the-fly” analyzer embedded on a web service. Apart from horizontal scaling considerations, threading could be considered to set several workers analyzing numerous documents at the same time.

It is expected that these next steps can be taken to extend the coverage of the research already performed.

Author Contributions: “Conceptualization, P.O.-C., C.M.-Á., M.C.-A., and M.I.D.-R.; methodology, P.O.-C., C.M.-Á., M.C.-A., and M.I.D.-R.; software, P.O.-C.; validation, P.O.-C., C.M.-Á., M.C.-A., and M.I.D.-R.; formal analysis P.O.-C.; investigation, P.O.-C., C.M.-Á., M.C.-A., and M.I.D.-R.; resources, P.O.-C.; data curation, P.O.-C.; writing—original draft preparation, P.O.-C., C.M.-Á., M.C.-A., and M.I.D.-R.; writing—review and editing, P.O.-C., C.M.-Á., M.C.-A., and M.I.D.-R. visualization, P.O.-C.; supervision, M.C.-A. and M.I.D.-R.; funding acquisition, M.C.-A. and M.I.D.-R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Xunta de Galicia, grant number ED431C 2017/50, ED481A-2018/275; Spanish Ministry of Economy and Competitiveness, grant number FFI2017-82752-P, and Autonomous Government of Galicia, grant number ED431C 2017/50.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wasserman, B.D.; Weseley, A.J. ¿ Qué? Quoi? Do languages with grammatical gender promote sexist attitudes? *Sex Roles* **2009**, *61*, 634. [[CrossRef](#)]
2. Meseguer, Á.G. *Es Sexista La Lengua Española? Una Investigación Sobre El Género Gramatical*; Editorial Paidós: Barcelona, Spain, 1996.
3. Publications Office of the European Union, EIGE (European Institute for Gender Equality). *Toolkit on Gender-sensitive Communication*; EIGE (European Institute for Gender Equality): Vilnius, Lithuania, 2018.
4. Kaufmann, C.; Bohner, G. Masculine generics and gender-aware alternatives in Spanish. In *Izgonzeit. Onlinezeitschrift Des Interdiszip. Zent. Für Geschlechterforschung (Izgj)*; University of Bielefeld: Bielefeld, Germany, 2014; Volume 1, pp. 8–17.
5. *Inclusive Use of Language, Guide for Authors*; Elsevier: Amsterdam, The Netherlands, 2020; Available online: <https://www.elsevier.com/journals/language-and-communication/0271-5309/guide-for-authors> (accessed on 8 April 2020).
6. Lakoff, R. Language and woman’s place. *Lang. Soc.* **1973**, *2*, 45–79. [[CrossRef](#)]
7. Mills, S. *Discourse*; Routledge: Abingdon-on-Thames, UK, 2004.
8. Cameron, D. *On Language and Sexual Politics*; Routledge: Abingdon-on-Thames, UK, 2012.
9. Fernández, Á.M.C. *Sexismo lingüístico. Análisis y Propuestas ante la Discriminación Sexual en el Lenguaje*; Narcea: Madrid, Spain, 1999.
10. Eckert, P.; McConnell-Ginet, S. Putting communities of practice in their place. *Gend. Lang.* **2007**, *1*, 27–37. [[CrossRef](#)]
11. Holmes, J.; Meyerhoff, M. Different voices, different views: An introduction to current research in language and gender. In *The Handbook of Language and Gender*; John Wiley & Sons: Hoboken, NJ, USA, 2003; pp. 1–17.
12. Prewitt-Freilino, J.L.; Caswell, T.A.; Laakso, E.K. The gendering of language: A comparison of gender equality in countries with gendered, natural gender, and genderless languages. *Sex Roles* **2012**, *66*, 268–281. [[CrossRef](#)]
13. Newman, M.L.; Groom, C.J.; Handelman, L.D.; Pennebaker, J.W. Gender differences in language use: An analysis of 14,000 text sample. *Discourse Process.* **2008**, *45*, 211–236. [[CrossRef](#)]
14. Foertsch, J.; Gernsbacher, M.A. In search of gender neutrality: Is singular they a cognitively efficient substitute for generic he? *Psychol. Sci.* **1997**, *8*, 106–111. [[CrossRef](#)] [[PubMed](#)]
15. Magner, T.F. Sexist and non-sexist usages in the English language. *Studia Romanica Et Anglica Zagrabienis: Revue; Publiée Par Les Sections Romane, Italienne Et Anglaise De La Faculté Des Lettres De l’Université De Zagreb. Transactions on Maritime Science* **2002**, *47*, 271–282.
16. Stout, J.G.; Dasgupta, N. When he doesn’t mean you: Gender-exclusive language as ostracism. *Pers. Soc. Psychol. Bull.* **2011**, *37*, 757–769. [[CrossRef](#)]
17. Sarrasin, O.; Gabriel, U.; Gygas, P. Sexism and attitudes toward gender-neutral language. *Swiss J. Psychol.* **2012**, *71*, 113–124. [[CrossRef](#)]
18. Verweken, D.; Hannover, B.; Wolter, I. Changing (S) expectations: How gender fair job descriptions impact children’s perceptions and interest regarding traditionally male occupations. *J. Vocat. Behav.* **2013**, *82*, 208–220. [[CrossRef](#)]

19. Gustafsson Sendén, M.; Bäck, E.A.; Lindqvist, A. Introducing a gender-neutral pronoun in a natural gender language: The influence of time on attitudes and behavior. *Front. Psychol.* **2015**, *6*, 893.
20. Chen, J.; Su, J. Differential sensitivity to the gender of a person by English and Chinese speakers. *J. Psycholinguist. Res.* **2011**, *40*, 195–203. [[CrossRef](#)] [[PubMed](#)]
21. Qiu, L.; Swaab, T.Y.; Chen, H.-C.; Wang, S. The role of gender information in pronoun resolution: Evidence from Chinese. *PLoS ONE* **2012**, *7*, e36156. [[CrossRef](#)] [[PubMed](#)]
22. Dong, Y.; Wen, Y.; Zeng, X.; Ji, Y. Exploring the cause of English pronoun gender errors by Chinese learners of English: Evidence from the self-paced reading paradigm. *J. Psycholinguist. Res.* **2015**, *44*, 733–747. [[CrossRef](#)] [[PubMed](#)]
23. Formanowicz, M.; Bedynska, S.; Cislak, A.; Braun, F.; Sczesny, S. Side effects of gender-fair language: How feminine job titles influence the evaluation of female applicants. *Eur. J. Soc. Psychol.* **2013**, *43*, 62–71.
24. Cacciari, C.; Carreiras, M.; Cionini, C.B. When words have two genders: Anaphor resolution for Italian functionally ambiguous words. *J. Mem. Lang.* **1997**, *37*, 517–532. [[CrossRef](#)]
25. Merkel, E.; Maass, A.; Frommelt, L. Shielding women against status loss: The masculine form and its alternatives in the Italian language. *J. Lang. Soc. Psychol.* **2012**, *31*, 311–320. [[CrossRef](#)]
26. Lévy, A.; Gygas, P.; Gabriel, U. Fostering the generic interpretation of grammatically masculine forms: When my aunt could be one of the mechanics. *J. Cogn. Psychol.* **2014**, *26*, 27–38. [[CrossRef](#)]
27. Baron, D.E. *Grammar and Gender*; Yale University Press: New Haven, CT, USA, 1986.
28. Gastil, J. Generic pronouns and sexist language: The oxymoronic character of masculine generics. *Sex Roles* **1990**, *23*, 629–643. [[CrossRef](#)]
29. Hamilton, M.C. Using masculine generics: Does generic he increase male bias in the user's imagery? *Sex Roles* **1988**, *19*, 785–799. [[CrossRef](#)]
30. Conkright, L.; Flannagan, D.; Dykes, J. Effects of pronoun type and gender role consistency on children's recall and interpretation of stories. *Sex Role* **2000**, *43*, 481–497. [[CrossRef](#)]
31. Ansara, Y.G.; Hegarty, P. Methodologies of misgendering: Recommendations for reducing cisgenderism in psychological research. *Fem. Psychol.* **2014**, *24*, 259–270. [[CrossRef](#)]
32. MEC, Ministerio de Educación y Ciencia. *Recomendaciones para el uso no sexista de la lengua*; MEC, Ministerio de Educación y Ciencia: Madrid, Spain, 1998.
33. Martínez, A.S. El sexismo ¿lingüístico? *Interlingüística* **2009**, *1*, 990–999.
34. Bosque, I. *Sexismo Lingüístico Y Visibilidad De La Mujer*; Real Academia Española: Madrid, Spain, 2012.
35. Sabater, C.P. Research on sexist language in EFL Literature: Towards a non-sexist approach. *Porta Linguarum.Revista Internacional De Didáctica De Las Lenguas Extranjeras* **2015**, *23*, 187–203.
36. Cabello Pino, M. Academias de la lengua española frente a guías de lenguaje no sexista: Un problema de delimitación de competencias. *Tonos Digital* **2019**, *37*, 1–30.
37. García Lopez, Á.; Morant, R. *Gramática Femenina*; Cátedra: Madrid, Spain, 1991.
38. Cabrera, M.J.C. "Acerca de la discriminación de la mujer y de los lingüistas en la sociedad. Reflexiones críticas". *Infoling* **2012**, *1*, 1–11.
39. Márquez Guerrero, M.S. Bases epistemológicas del debate sobre el sexismo lingüístico. *Arbor: Ciencia, Pensamiento Y Cultura* **2016**, *192*, a307. [[CrossRef](#)]
40. Mostafa, M.M. More than words: Social networks' text mining for consumer brand sentiments. *Expert Syst. Appl.* **2013**, *40*, 4241–4251. [[CrossRef](#)]
41. Nassirtoussi, A.K.; Aghabozorgi, S.; Ying Wah, T.; Ngo, D.C.L. Text mining for market prediction: A systematic review. *Expert Syst. Appl.* **2014**, *41*, 7653–7670. [[CrossRef](#)]
42. Gonzalez, G.H.; Tahsin, T.; Goodale, B.C.; Greene, A.C.; Greene, C.S. Recent advances and emerging applications in text and data mining for biomedical discovery. *Brief. Bioinform.* **2015**, *17*, 33–42. [[CrossRef](#)]
43. Kumar, B.S.; Ravi, V. A survey of the applications of text mining in financial domain. *Knowl. Based Syst.* **2016**, *114*, 128–147. [[CrossRef](#)]
44. Haddoud, M.; Mokhtari, A.; Lecroq, T.; Abdeddaim, S. Combining supervised term-weighting metrics for SVM text classification with extended term representation. *Knowl. Inf. Syst.* **2016**, *49*, 909–931. [[CrossRef](#)]
45. Pratama, B.Y.; Sarno, R. Personality classification based on twitter text using naive bayes, KNN and SVM. In Proceedings of the 2015 International Conference on Data and Software Engineering (ICoDSE), Yogyakarta, Indonesia, 25–26 November 2015; pp. 170–174.

46. Lin, Y.; Wang, J. Research on text classification based on SVM-KNN. In Proceedings of the 2014 IEEE 5th International Conference on Software Engineering and Service Science, Beijing, China, 27 June 2014; pp. 842–844.
47. Yin, C.; Xiang, J.; Zhang, H.; Wang, J.; Yin, Z.; Kim, J.-U. A new SVM method for short text classification based on semi-supervised learning. In Proceedings of the 2015 4th International Conference on Advanced Information Technology and Sensor Application (AITS), Harbin, China, 21–23 August 2015; pp. 100–103.
48. Trstenjak, B.; Mikac, S.; Donko, D. KNN with TF-IDF based framework for text categorization. *Procedia Eng.* **2014**, *69*, 1356–1364. [[CrossRef](#)]
49. Jiang, L.; Li, C.; Wang, S.; Zhang, L. Deep feature weighting for naive Bayes and its application to text classification. *Eng. Appl. Artif. Intell.* **2016**, *52*, 26–39. [[CrossRef](#)]
50. Shimodaira, H. Text classification using naive Bayes. *Learn. Data Note* **2014**, *7*, 1–9.
51. Wang, S.; Jiang, L.; Li, C. Adapting naive Bayes tree for text classification. *Knowl. Inf. Syst.* **2015**, *44*, 77–89. [[CrossRef](#)]
52. Agarwal, B.; Mittal, N. Text classification using machine learning methods—a survey. In Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), Jaipur, India, 28–30 December 2012; Springer: New Dehli, India, 2014; pp. 701–709.
53. Brindha, S.; Prabha, K.; Sukumaran, S. A survey on classification techniques for text mining. In Proceedings of the 2016 3rd International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 22–23 January 2016; pp. 1–5.
54. Allahyari, M.; Pouriyeh, S.; Assefi, M.; Safaei, S.; Trippe, E.D.; Gutierrez, J.B.; Kochut, K. A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv* **2017**, arXiv:1707.02919.
55. Jindal, R.; Malhotra, R.; Jain, A. Techniques for text classification: Literature review and current trends. *Webology* **2015**, *12*, 6–12.
56. Zhang, X.; Zhao, J.; LeCun, Y. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2015; pp. 649–657.
57. Davidson, T.; Warmsley, D.; Macy, K.; Weber, I. Automated hate speech detection and the problem of offensive language. In Proceedings of the Eleventh International Aaai Conference on Web and Social Media, Montreal, QC, Canada, 15–18 May 2017.
58. Ashcroft, M.; Fisher, A.; Kaati, L.; Omer, E. Detecting jihadist messages on twitter. In Proceedings of the 2015 European Intelligence and Security Informatics Conference, Manchester, UK, 7–9 September 2015; pp. 161–164.
59. Dias, D.S.; Welikala, M.D.; Dias, N.G. Identifying racist social media comments in sinhala language using text analytics models with machine learning. In Proceedings of the 2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer), Colombo, Sri Lanka, 26–29 September 2018; pp. 1–6.
60. Sharifirad, S.; Jafarpour, B.; Matwin, S. Boosting text classification performance on sexist tweets by text augmentation and text generation using a combination of knowledge graphs. In Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), Brussels, Belgium, 31 October 2018; pp. 107–114.
61. Pitsilis, G.K.; Ramampiaro, H.; Langseth, H. Detecting offensive language in tweets using deep learning. *arXiv* **2018**, arXiv:1801.04433.
62. Trivedi, M.; Sharma, S.; Soni, N.; Nair, S. Comparison of text classification algorithms. *IJERT* **2015**, *4*, 11–23.
63. Jain, A.; Mandowara, J. Text classification by combining text classifiers to improve the efficiency of classification. *Int. J. Comput. Appl.* **2016**, *6*, 655–660.
64. Dumais, S. Using SVMs for text categorization. *IEEE Intell. Syst.* **1998**, *13*, 21–23.
65. Basu, A.; Walters, C.; Shepherd, M. Support vector machines for text categorization. In Proceedings of the 36th Annual Hawaii International Conference on System Sciences, Big Island, HI, USA, 6–9 January 2003; p. 7.
66. Gonen, H.; Goldberg, Y. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv* **2019**, arXiv:1903.03862.
67. Dixon, L.; Li, J.; Sorensen, J.; Thain, N.; Vasserman, L. Measuring and mitigating unintended bias in text classification. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, New Orleans, LA, USA, 2–3 February 2018; pp. 67–73.

68. Nozza, D.; Volpetti, C.; Fersini, E. Unintended bias in misogyny detection. In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, Thessaloniki, Greece, 14–17 October 2019; pp. 149–155.
69. Bolukbasi, T.; Chang, K.-W.; Zou, J.; Saligrama, V.; Kalai, A. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2016; pp. 4349–4357.
70. Badjatiya, P.; Gupta, M.; Varma, V. Stereotypical bias removal for hate speech detection task using knowledge-based generalizations. In Proceedings of the The World Wide Web Conference, San Francisco, CA, USA, 14–17 May 2019; pp. 49–59.
71. Basile, V.; Bosco, C.; Fersini, E.; Nozza, D.; Patti, V.; Rangel, F.; Rosso, P.; Sanguinetti, M. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In Proceedings of the 13th International Workshop on Semantic Evaluation, Minneapolis, MN, USA, 26 June 2019; pp. 54–63.
72. Zhou, P.; Shi, W.; Zhao, J.; Huang, K.-H.; Chen, M.; Cotterell, R.; Chang, K.-W. Examining gender bias in languages with grammatical gender. *arXiv* **2019**, arXiv:1909.02224.
73. Universidad de La Rioja. Buscador de tesis doctorales. Available online: <https://dialnet.unirioja.es/institucion/unirioja/tesis> (accessed on 15 December 2019).
74. Archana, S.; Elangovan, K. Survey of classification techniques in data mining. *Int. J. Comput. Sci. Mob. Appl.* **2014**, *2*, 65–71.
75. Patra, A.; Singh, D. A survey report on text classification with different term weighing methods and comparison between classification algorithms. *Int. J. Comput. Appl.* **2013**, *75*, 2–5. [[CrossRef](#)]
76. Santafe, G.; Inza, I.; Lozano, J.A. Dealing with the evaluation of supervised classification algorithms. *Artif. Intell. Rev* **2015**, *44*, 467–508.
77. Blini, L. Usos inclusivos de género en el castellano legislativo de la Unión Europea y de España. In *Gender in Legislative Languages: From EU to National Law in English, French, German, Italian and Spanish*; Frank et Timme: Berlin, Germany, 2019; Volume 144, p. 183.
78. Cortés, L.; de la Paz, M. *Comunicación Política Local con Perspectiva de Género: Análisis y Propuesta de Mejora del Lenguaje Inclusivo Administrativo con Perspectiva de Género en la Red Social de Facebook de los Ayuntamientos de Arjona (Jaén) y Bollullos de la Mitación (Sevilla)*; Publicaciones de Universidad de Sevilla: Sevilla, Spain, 2018.
79. Fundación, O. *Guía para un uso no sexista del lenguaje: Incluye una mirada especial al empleo ya la discapacidad*; Publicaciones Fundación ONCE: Madrid, Spain, 2019.
80. Marçal, H.; Kelso, F.; Nogués, M. *Guía para el uso no sexista del lenguaje en la Universidad Autónoma de Barcelona*; Servicio de Publicaciones de la UAB: Barcelona, Spain, 2011.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).