TOP
https://doi.org/10.1007/s11750-024-00671-1

ORIGINAL PAPER

Adapting support vector optimisation algorithms to textual gender classification

Javier Gomez[1] · Cesar Alfaro[1] · Felipe Ortega[1] · Javier M. Moguerza[1] · Maria Jesus Algar[1] · Raul Moreno[2]

Received: 31 March 2023 / Accepted: 7 March 2024

Abstract
In this paper, we focus on the problem of determining the gender of the person described in a biographical text. Since support vector machine classifiers are well suited for text classification tasks, we present a new stopping criterion for support vector optimisation algorithms tailored to this problem. This new approach exploits the geometric properties of the vector representation of such content. An experiment on a set of English and Spanish biographical articles retrieved from Wikipedia illustrates this approach and compares it to other machine learning classification algorithms. The proposed method allows real-time classification algorithm training. Moreover, these results confirm the advantage of leveraging additional gender information in strongly inflected languages, like Spanish, for this task.

Keywords Support vector machines · Machine learning · Nonlinear optimisation · Text mining · Gender identification

Mathematics Subject Classification 68U15 · 68T50

1 Introduction

The availability of large and rich textual datasets, mainly created from web content, has fostered data mining and machine learning algorithms and methods at an unprecedented pace. As a result, machine learning for text analysis has become a very active research area Aggarwal (2018). Nowadays, *text mining*Srivastava and Sahami (2009) and *Natural Language Processing* (NLP) Eisenstein (2019) constitute the basis for many studies and implementations.

The Wikidata project Vrandečić and Krötzsch (2014) is an example of a freely available and structured data source nurtured by information extracted from

Javier Gomez, Cesar Alfaro, Felipe Ortega, Javier M. Moguerza, Maria Jesus Algar, and Raul Moreno have contributed equally to this work.

Extended author information available on the last page of the article

Published online: 13 April 2024

Wikipedia articles in different languages. Data gathered from Wikipedia articles undergo a semantic annotation process to expand the Wikidata knowledge base, following a graph-structured data model to integrate new applications.

One such application area is gender identification in text. The goal is to assign a gender label (e.g., "female" or "male") to a text block based on linguistic properties and other descriptive features. In general, previous research has focused on identifying the gender and profile of the person who writes the text. Examples include gender identification of weblog authors Yan and Yan (2006) and users in microblogging platforms Mukherjee and Bala (2017) or in Customer Relationship Management (CRM) systems Amado et al. (2018).

Building new Wikidata entries relies on existing hints and structured data in Wikipedia articles to complete each topic. For instance, if we want to retrieve structured data about "Gabriel García Marquez", we can search in Wikidata for the corresponding entry.[1] Then, we can explore several fields with descriptive features about this person, including the "sex or gender" property.[2] In this example, the attributed value for this property is "male". At the time of this writing, the accuracy of this information relies on up to six references to different Wikipedia articles that support the attribution of the "male" value to this field based on structured data. However, as an open project for collaborative knowledge creation, Wikipedia has frequently suffered vandalism and content attacks Adler et al. (2011), Geiger and Ribes (2010). If malicious agents alter original structured data in Wikipedia articles, detecting errors without manual inspection would be a daunting task, even more so given the large size of the Wikidata topic database. Moreover, if structured information about gender is not present in a given Wikipedia article, it would not be possible to assign a value directly to the "sex or gender" property for its corresponding Wikidata entry. Instead, labelling the gender information based on an automated analysis of unstructured textual data in the body of Wikipedia articles could improve resilience against these issues. This example can be easily extrapolated to any other information system that must automatically tag the content of some text block with a gender label.

We focus on a specific case, namely, the problem of automatically determining the gender of the person described in a biographical text. In the remainder of this work, for the sake of space, we will refer to this problem as gender identification. The lack of studies that include languages other than English hinders the development of novel procedures in text mining to tackle this problem better. Instead of focusing on English exclusively, which provides fewer gender elements, we examine the use of Spanish text to capture morphemes that carry gender information and are more prevalent in that language. To achieve this, we bypass usual normalisation steps in traditional text mining, such as stemming and lemmatisation Eisenstein (2019), that discard gender information in linguistic terms. Following this approach, we show that it is possible to assign a gender value to textual content more accurately.

When processing text using machine learning techniques, it is common to use vector representations, characterised by high dimensionality and sparsity, since they

[1] https://www.wikidata.org/wiki/Q5878.

[2] https://www.wikidata.org/wiki/Property:P21.

contain numerous zero-valued components. In this context of high-dimensional sparse representations, Support Vector Machines (SVM) Moguerza and Muñoz (2006) are computationally adequate classifiers Joachims (2002). However, adapting traditional optimisation problem-solving techniques for SVM classification Joachims (1999) applied to gender identification does not guarantee rapid convergence when the iterations approach the optimal solution of the underlying optimisation problem.

In this paper, within a support vector framework, we take advantage of the geometric properties of the text gender identification task to modify the stopping criterion of the underlying optimisation algorithm by substituting the classical Karush–Kuhn–Tucker criterion Feng and Li (2018) with our proposal. It incorporates information (the error rate) that is not directly mathematically related to the criterion that the classical optimisation SVM algorithm tries to fulfil, as it can potentially adapt the nature of the algorithm to the geometric problem to resolve. In the following, we will refer to this modified SVM as Geometrical SVM (GSVM). To evaluate the performance of this new approach, we compare our proposal to other supervised learning methods Witten et al. (2017) for automated gender identification within biographical texts of literary authors extracted from Wikipedia. We assess if avoiding the common step of stemming in text mining, which eliminates gender suffixes in relevant words for text analysis, improves performance in gender identification. We show the effectiveness of the GSVM approach for the acceleration of the algorithm training phase and that this method does not significantly affect the procedure performance.

The rest of this paper is organised as follows. Section 2 reviews related research work and identifies limitations of existing approaches for gender identification in textual data. Section 3 briefly describes the foundations of SVM. Section 4 develops the proposed new stopping criterion within support vector optimisation algorithms. Then, Sect. 5 shows the problem statement and the experimental setup to validate the proposed method. Section 6 presents the numerical results. Section 7 discusses additional aspects and some limitations of our approach and lays out possible lines for future work. Finally, we summarise the main conclusions from this research in Sect. 8.

2 Related work on gender identification

Here, we present a summary of previous research work on gender identification, including several methodological aspects that exert a direct influence on this setting.

2.1 Machine learning and text mining

Identifying and extracting useful information and patterns from textual data are a very active research area in machine learning and artificial intelligence Aggarwal (2018). Information Retrieval Baeza-Yates and Ribeiro-Neto (2011), Natural Language Processing (NLP) Eisenstein (2019), Jurafsky and Martin (2009), and text

mining Feldman and Sanger (2006), Srivastava and Sahami (2009), Berry and Kogan (2010) are related knowledge areas providing methods and tools for this task.

Although much of this previous work has considered English textual data, some studies have also focused on methodological aspects of text analysis in other languages, like in this case. In particular, Hedlund et al. (2001) study the use of Swedish for cross-language information retrieval. They highlight specific traits that could be exploited in certain applications, for example, morphological features, such as inflexion, derivation, and gender. In our approach, we attempt to leverage similar morphological elements in Spanish to improve gender identification in text.

Besides, our method introduces some changes in the conventional pipeline for text data preparation to retain these valuable morphological elements for gender identification. Previous research confirms the importance of selecting appropriate combinations of pre-processing tasks to improve the accuracy of machine learning algorithms for classification Uysal and Gunal (2014). That is even more relevant in the analysis of languages different from English.

2.2 Identifying the gender of text authors

Gender identification in textual data has attracted significant interest in previous research. One of the most frequent applications have been detecting the gender of the person who writes the text, including microblogging users Mukherjee and Bala (2017), Huang et al. (2014), authors of e-mail messages Corney et al. (2002), chat messages Kucukyilmaz et al. (2006) and weblog posts Yan and Yan (2006), contributors in peer-production online projects Vasilescu et al. (2014), Lin and Serebrenik (2016), Terrell et al. (2017), Das et al. (2019), or gender detection of feedback authors in CRM platforms Lau et al. (2005), Amado et al. (2018). Profiling text authors has also been addressed in multilingual settings Kocher and Savoy (2017),Fatima et al. (2017), López-Santillán et al. (2020), confirming the advantages of content-based methods for this task. Other methodological approaches in this context include the use of graph analysis Kretschmer and Aguillo (2005), Rangel and Rosso (2016) and Part Of Speech (POS) tagging Fourkioti et al. (2019).

In Krüger and Hermann (2019), the authors examine 215 previous research works published between 2017 and 2019 on the gender identification of text authors. According to this review, the best experiment in previous literature reports an accuracy of 93.4% on cases extracted from Facebook Posts Markov et al. (2017). It is worth mentioning that all papers considered in this survey follow standard machine learning procedures to prepare textual input, including tokenisation and text normalisation Eisenstein (2019) (such as lemmatisation or stemming). As we explain in Sect. 5, in our approach, we introduce changes in standard procedures to improve the performance in gender identification as suggested in previous research Uysal and Gunal (2014). In addition, a common limitation raised in this comparison is that all previous studies only consider a binary gender target (we will discuss again this matter in Sect. 8).

Furthermore, according to Krüger and Hermann (2019), there is ample variability in gender identification accuracy across different languages. A closer inspection of

results from featured works in this review reveals that experiments with languages providing richer gender information, such as Spanish and Portuguese, outperform the accuracy in English. In our method, we seek to confirm if it is possible to leverage specific features in some languages for gender identification purposes.

2.3 Text gender identification

Text gender identification, that is, deciding the gender of a person described in a text, has received comparatively less attention in previous studies. Nonetheless, there are applications for targeted advertising Jansen et al. (2013), evaluating gender differences in online labour markets Foong et al. (2018) and named entity recognition Cho et al. (2013). In contrast with prior approaches, predominantly based on database services Wais (2016), Santamaría and Mihaljević (2018), more recent work by Das and Paik (2021) shows the utility of applying machine learning algorithms that analyse contextual information. We propose to combine machine learning algorithms with the retention of morphological elements that carry gender information in some languages to further improve the accuracy of gender tagging a person described in text.

3 Background on SVM optimisation

SVM are algorithms whose performance is based on the use of kernels. A kernel $K(x, y)$ is a real-valued function $K : X \times X \to \mathbb{R}$ that acts as a dot product in a real vector space Z. To this aim, there exists a function

$$\Phi : X \to Z, \tag{1}$$

such that $K(x, y) = \Phi(x)^T \Phi(y)$. X and Z are, respectively, known as input and feature spaces.

SVM belong to the type of algorithms based on regularisation theory Moguerza and Muñoz (2006). These methods allow the construction of classification functions by solving an optimisation problem of the form (Tikhonov and Arsenin 1977)

$$\min_{f \in H_K} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) + \mu \|f\|_K^2, \tag{2}$$

where $\mu > 0$; H_K denotes the reproducing kernel Hilbert space (RKHS) associated with the kernel K; $\|f\|_K$ is the norm of f in the RKHS; x_i are the n sampled data points; $y_i \in \{-1, +1\}$ indicates the two possible classes of x_i; and, finally, $L(y_i, f(x_i))$ is an error function. In this context, $f(x) = 0$ is a decision surface in H_K. The typical SVM approach uses the specific error function L, called hinge loss, defined as

$$L(y_i, f(x_i)) = (1 - y_i f(x_i))_+, \tag{3}$$

with $(x)_+ = \max(x, 0)$.

In problem (2), μ helps to establish a compromise between the fit of solution f to the data, quantified by L, and the complexity of function f, quantified by $\|f\|_K$.

It is immediate to show that $\|f\|_K^2 = \|w\|^2$, where $w = \sum_i^n \alpha_i \Phi(x_i)$ and Φ is the mapping defined in (1). Therefore, problem (2) can be reformulated as

$$\min_{w,b} \frac{1}{n} \sum_{i=1}^{n} (1 - y_i(w^T \Phi(x_i) + b))_+ + \mu \|w\|^2. \tag{4}$$

Problem (4) is equivalent to solving the following optimisation dual problem Moguerza and Muñoz (2006):

$$\begin{aligned} \min_{\beta} \quad & \frac{1}{2} \boldsymbol{\beta}^T Q \boldsymbol{\beta} - \boldsymbol{e}^T \boldsymbol{\beta} \\ \text{s.t.} \quad & \boldsymbol{y}^T \boldsymbol{\beta} = 0, \\ & \mathbf{0} \leq \boldsymbol{\beta} \leq C\boldsymbol{e}, \end{aligned} \tag{5}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)^T$, $\boldsymbol{y} = (y_1, \ldots, y_n)^T$, $\mathbf{0}$ is a vector of all zeros, $\boldsymbol{e}$ is a vector of all ones, Q is a positive definite $n \times n$ symmetric matrix with $Q_{ij} = y_i y_j K(x_i, x_j)$, and $C = \frac{1}{2\mu n}$ is a constant. This problem is convex and quadratic and, therefore, every local minimum is a global minimum.

For the sake of simplicity, let us define $g(\boldsymbol{\beta})$ as the gradient of the objective function of problem (5), that is, $g(\boldsymbol{\beta}) = Q\boldsymbol{\beta} - \boldsymbol{e}$.

A vector $\boldsymbol{\beta}$ is a stationary point of (5) if and only if there is a number d and two non negative vectors $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$, such that

$$\begin{aligned} g(\boldsymbol{\beta}) + dy &= \boldsymbol{\lambda} - \boldsymbol{\mu}, \\ \boldsymbol{\lambda_i} \boldsymbol{\beta_i} &= 0, i = 1 \ldots n, \\ \boldsymbol{\mu_i}(C - \boldsymbol{\beta_i}) &= 0, i = 1 \ldots n, \\ \boldsymbol{\lambda_i}, \boldsymbol{\mu_i} &\geq 0, i = 1 \ldots n. \end{aligned} \tag{6}$$

It can be shown, Chen et al. (2006), that a vector $\boldsymbol{\beta}$ such that $0 \leq \boldsymbol{\beta_i} \leq C$ satisfies conditions (6) if and only if

$$\begin{aligned} -y_i g(\boldsymbol{\beta})_i \leq d, \quad & \forall i \in I_{up}(\boldsymbol{\beta}), \\ -y_i g(\boldsymbol{\beta})_i \geq d, \quad & \forall i \in I_{low}(\boldsymbol{\beta}), \end{aligned} \tag{7}$$

where

$$\begin{aligned} I_{up}(\boldsymbol{\beta}) &\equiv \{t \mid \beta_t < C, y_t = 1 \text{ or } \beta_t > 0, y_t = -1\}, \text{and} \\ I_{low}(\boldsymbol{\beta}) &\equiv \{t \mid \beta_t < C, y_t = -1 \text{ or } \beta_t > 0, y_t = 1\}. \end{aligned} \tag{8}$$

From (7), it holds that $m(\boldsymbol{\beta}) \leq M(\boldsymbol{\beta})$, where

$$\begin{aligned} m(\boldsymbol{\beta}) &\equiv \max_{i \in I_{up}(\boldsymbol{\beta})} -y_i g(\boldsymbol{\beta})_i, \\ M(\boldsymbol{\beta}) &\equiv \min_{i \in I_{low}(\boldsymbol{\beta})} -y_i g(\boldsymbol{\beta})_i, \end{aligned} \tag{9}$$

In the SVM literature, problem (5) is solved using the so-called Sequential Minimal Optimisation (SMO) type algorithms (Joachims 1999; Platt 1998). These algorithms are essentially Newton-type quadratic methods that, to make the problem computationally tractable, consider only a subset of variables in each iteration, the so-called working set, instead of working with the entire matrix $\boldsymbol{Q}$.

Theorem 1 *Let* $\{\boldsymbol{\beta}^k\}$ *be the infinite sequence generated by an SMO-type method for problem* (5). *Then, if* Q *is a positive definite matrix, the limit point of* $\{\boldsymbol{\beta}^k\}$ *is the unique and global minimum of problem* (5).

Proof See Chen et al. (2006). □

As a consequence, the following corollary holds.

Corollary 3.1 *If* $\{\boldsymbol{\beta}^k\}$ *is an infinite sequence, then the following two limits exist and are equal*:

$$\lim_{k\to\infty} m(\boldsymbol{\beta}^k) = \lim_{k\to\infty} M(\boldsymbol{\beta}^k). \tag{10}$$

Considering the definition of the sets I_{up} and I_{low} from Eq. (8), and in particular the $\boldsymbol{\beta}^k$ involved within each set (within bounds), for k large enough and a small tolerance $\epsilon > 0$, the following condition holds:

$$m(\boldsymbol{\beta}^k) - M(\boldsymbol{\beta}^k) \leq \epsilon, \tag{11}$$

that is

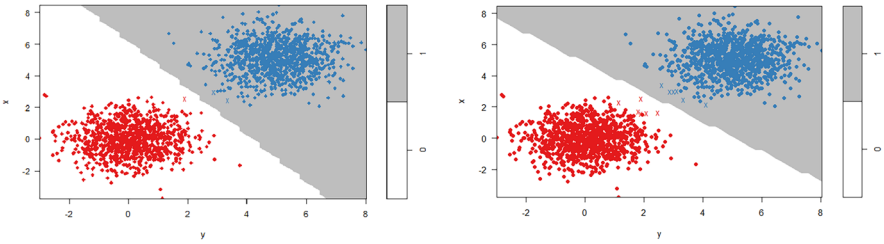
$$\lim_{k\to\infty} |m(\boldsymbol{\beta}^k) - M(\boldsymbol{\beta}^k)| = 0.$$

Based on these results, SMO-type algorithms implement the following stopping criterion:

$$|m(\boldsymbol{\beta}^k) - M(\boldsymbol{\beta}^k)| \leq \epsilon. \tag{12}$$

4 Geometrical stopping criterion

The geometry of the SVM decision function linked to the training data is quantifiable within each iteration by measuring or estimating the error rate. In this sense, when such an error rate stabilises, the decision function fulfils the requirements for successfully classifying the data. Figure 1 shows an example with two 2-dimensional Gaussian clouds. Running an SVM up to iteration 2 produces the separating surface represented in Fig. 1a, whereas Fig. 1b exhibits the separating surface calculated up to iteration 18. The decision function in Fig. 1a provides the same empirical error as the decision function in Fig. 1b, which required



(a) SVM run until iteration 2 (b) SVM run until iteration 18

Fig. 1 Two separating surfaces calculated with SVM: (**a**) separating surface obtained after 2 iterations; (**b**) separating surface calculated after 18 iterations

more iterations and, therefore, more training time for its construction. In contrast, building the decision function in Fig. 1a needed fewer training iterations.

Based on these foundations, let us define ε_k as the error rate for the training set at iteration k and $\tau > 0$ as a real value acting as a tolerance. According to (3)

$$\epsilon_k = \frac{1}{n}\sum_{i=1}^{n}(1 - y_i f(x_i))_+. \tag{13}$$

Let $\delta_k = |\varepsilon_k - \varepsilon_{k-5}|$. The method will stop when it holds that

$$(\delta_k \leq \tau) \wedge (\delta_{k-5} \leq \tau) \wedge (\delta_{k-15} \leq \tau) \wedge (|M(\boldsymbol{\beta}^k) - m(\boldsymbol{\beta}^k)| \leq 100\epsilon), \tag{14}$$

where the symbol "$\wedge$" denotes the logical operator "and", that is, the four inequalities within the criterion must hold simultaneously. By measuring inequality $\delta_k \leq \tau$ every five iterations and relaxing the classical criterion, we can avoid undesired error peaks caused by randomness. The following theorem demonstrates that there is a direct relationship between the stopping criteria (12) and (14).

Theorem 2 *Given $\tau > 0$ and $\epsilon > 0$, there exists an iteration number $k_{\tau,\epsilon}$, such that $\delta_k \leq \tau$ and $|m(\boldsymbol{\beta}^k) - M(\boldsymbol{\beta}^k)| \leq \epsilon, \forall k > k_{\tau,\epsilon}$.*

Proof By Corollary 3.1, it holds that

$$\lim_{k\to\infty} |m(\boldsymbol{\beta}^k) - M(\boldsymbol{\beta}^k)| = 0.$$

Thus, there exists an iteration number k_ϵ, such that $\forall k > k_\epsilon$, it holds that

$$|m(\boldsymbol{\beta}^k) - M(\boldsymbol{\beta}^k)| \leq \epsilon.$$

Let the vector $\boldsymbol{\beta}^*$ denote the solution to problem (5), where x_i such that $\beta_i^* > 0$ are the so-called support vectors. It is well known that the function f^*, which determines the decision surface $f^*(x) = 0$, takes the form

$$f^*(x) = \sum_{i=1}^{n} \beta_i^* y_i K(x_i, x) + b^*, \tag{15}$$

with

$$b^* = -\frac{\sum_{i=1}^{n} \beta_i^* y_i K(x_i, x^+)}{2} + \frac{\sum_{i=1}^{n} \beta_i^* y_i K(x_i, x^-)}{2},$$

where x^+ and x^- are two support vectors in classes +1 and -1, respectively, such that their associated Lagrange multipliers β^+ and β^- hold that $0 < \beta^+ < C$ and $0 < \beta^- < C$.

Let us consider the decision function determined by problem (5), at iteration k

$$f^k(x) = \sum_{i=1}^{n} \beta_i^k y_i K(x_i, x) + b^k.$$

It is straightforward to show that, $\forall \gamma > 0$, there exists an iteration number k_γ, such that $\forall k > k_\gamma$

$$|f^k(x) - f^*(x)| \leq \gamma. \tag{16}$$

This is due to Theorem 1, which guarantees that

$$\lim_{k \to \infty} \boldsymbol{\beta}^k = \boldsymbol{\beta}^*,$$

and, hence, that $\forall \nu > 0$, there exists an iteration number k_ν such that, $\forall k > k_\nu$, it holds that

$$|\boldsymbol{\beta}^k - \boldsymbol{\beta}^*| \leq \nu.$$

Taking ν small enough, (16) holds for $k_\gamma = k_\nu$. Given $\tau > 0$, taking γ small enough in (16), $\forall k > k_\gamma$, considering definition (13), it holds that

$$\delta_k = |\varepsilon_k - \varepsilon_{k-5}| \leq \tau,$$

and the theorem holds for $k_{\tau,\epsilon} = \max\{k_\epsilon, k_\nu\}$. □

Since criterion (14) is based on the error rate, it is intuitive to fix a value for τ. We can set it to, for example, $\tau = 0.0005$, since an error rate of up to 0.05% can be considered significantly low. This intuition does not exist for criterion (12), because the magnitude of the $g(\boldsymbol{\beta})$ cannot be estimated in advance. In Sect. 5, we demonstrate that, for the particular case of text gender identification, we can find suitable generalising decision surfaces that fulfil criterion (14) in significantly fewer iterations than required to meet the commonly used theoretical criterion (12). The key point is the high-dimensional setting in which the text is represented, which allows reaching a good empirical generalisation quickly. Therefore, from an empirical point of view, it is expected that GSVM inherits the generalisation properties of SVM.

5 Experiments

In this section, we describe in detail the gender identification problem and the experimental setup.

5.1 Problem statement

We set out the problem of gender identification in textual data as a supervised learning classification task. Given a text block whose content can be tagged with a gender label associated with the person described in the text, the algorithm must automatically infer such a label. In this case, we consider two possible output labels, $L = \{\text{“female”}, \text{“male”}\}$, since our experimental dataset only contains instances of these two gender classes, as described in Sect. 5.3. Hence, we restrict our choice of algorithms to those geared towards binary classification. The same approach could be extended to consider additional output labels, reformulating the problem as a one-versus-all or a multi-label classification task.

Our framework differs from previous approaches in two main aspects:

- While prior works follow conventional text-mining procedures, such as stemming, we intentionally avoid this step to retain the gender suffix in terms and explore its impact on gender identification for text in strongly inflected languages like Spanish.
- Instead of tagging a sequence of named entities in a text, we tackle the problem of assigning a gender label to a block of text or a complete document that describes a person. That is a relevant problem for many applications, including automated annotation of entries in semantic databases.

Many languages are inflected, meaning that words can change their form to reflect grammatical information, such as number, tense, or gender Hedlund et al. (2001). In Spanish, words ending in “-o” or a consonant denote masculine gender, whereas those ending in “-a” are primarily feminine. Stemming attempts to remove the differences between inflected forms of a word to reduce each word to its root form. This pre-processing has been the standard approach in text mining and Information Retrieval to improve the precision in identifying key terms in a given text.

Nevertheless, as mentioned above, in Spanish, the suffix of many words can be a handy and direct indicator of gender. For this reason, it seems reasonable to assume that if we do not apply stemming and retain gender suffixes instead, we can improve accuracy in gender identification. Thus, we propose the following approach for gender identification in textual data:

1. Start by applying standard procedures in text mining to prepare the raw input text, including tokenisation and removal of stopwords (additional implementation details are provided in Sect. 5).

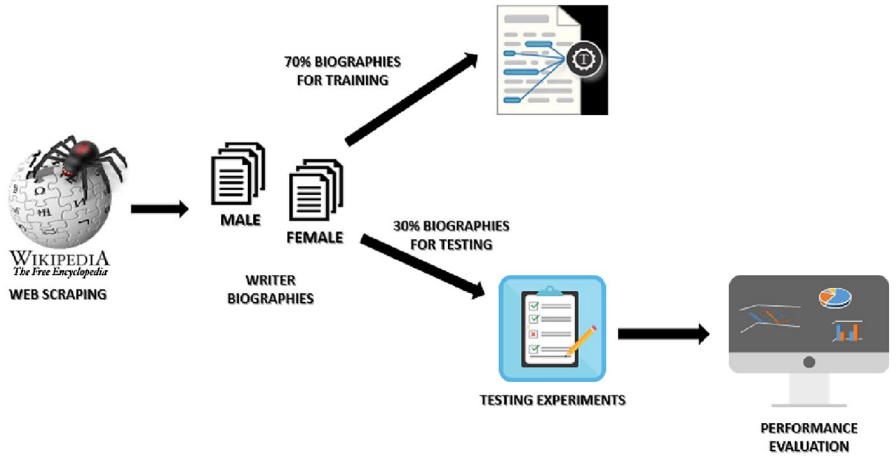


Fig. 2 Description of the proposed framework to test alternative supervised learning methods

2. Before creating the vectorised representation for each document, conventional text-mining procedures suggest performing some text normalisation procedure Sproat et al. (2001), such as stemming, to eliminate inflectional affixes, provide a typical representation of similar words and reduce the vocabulary size. On the contrary, we propose suppressing this text normalisation step, retaining instead inflectional suffixes that carry extra information about gender in specific languages.
3. Then, we resume the standard pipeline for document preparation, generating a vectorized representation of each document and building a term-document matrix (again, further details are provided in Sect. 5).

This strategy applies to preparing textual data that will train a machine learning algorithm for gender identification. To show the applicability of this procedure in a real setting, we have designed a series of validation experiments, described in the next section.

5.2 Experimental setup

The main objective is to evaluate the capacity of our proposed framework for automated identification of gender associated with real-world text documents. For this purpose, we retrieve biographical articles describing literary authors from the English and Spanish versions of Wikipedia, whose data feed other semantic knowledge-based systems like Wikidata. Therefore, two different datasets were created, one for each language. We have designed our dataset's construction to ensure equal representation of the two labels in biographical entries chosen for our experiments in English and Spanish.

Implementing our procedure on input text in English and Spanish, we aim to confirm if avoiding stemming has a positive impact on the performance of our gender

Table 1 Total and average number describing the English and Spanish datasets

Dataset	Descriptive statistics	Value
English	Total number of biographies of female authors	500
	Avg. number of distinct terms per female writer biography	50.86
	Avg. number of distinct terms per female writer biography (stemming)	38.11
	Total number of biographies of male authors	500
	Avg. number of distinct terms per biography of male writer	46.2
	Avg. number of distinct terms per male writer biography (stemming)	35.62
Spanish	Total number of biographies of female authors	416
	Avg. number of distinct terms per female writer biography	50.03
	Avg. number of distinct terms per female writer biography (stemming)	31.44
	Total number of biographies of male authors	416
	Avg. number of distinct terms per male writer biography	51.39
	Avg. number of distinct terms per male writer biography (stemming)	32.68

identification classifier. Our curated dataset consisted of biographies of female and male authors. This choice allows us to frame the experiment as a binary classification problem for which several standard machine learning algorithms exist. Figure 2 shows an overview of the experimental procedure to assess each supervised learning method.

Each dataset of biographies in English and Spanish is divided into training and testing subsets. Specifically, 70% of the biographies in each language are used for training, whereas the remaining 30% are held out as a testing set to assess the performance of supervised learning algorithms. The following section provides additional details about retrieving and preparing both datasets.

5.3 Datasets

To begin with, we retrieve biographies of literary authors from Wikipedia, crawling the web interface directly using the WikipediR package Keyes and Tilbert (2017), available for the R statistical environment R Core Team (2022). This package can obtain a list of pages, subcategories, page content, and other information about a specified category.

The English dataset Gomez et al. (2021a), publicly available,[3] consists of 1000 biographies about writers created in the English Wikipedia. These articles are equally partitioned into a female and a male set. Female biographies have been extracted from the category "19th-century_women_writers", whereas the first 500 pages of the "19th-century_male_writers" category have been obtained for male biographies.

The Spanish dataset Gomez et al. (2021b), also publicly available,[4] comprises 832 biographies from the Spanish Wikipedia. Female biographies have been extracted

[3] https://doi.org/10.6084/m9.figshare.13551467.
[4] https://doi.org/10.6084/m9.figshare.13551437.

Table 2 Example of term-biography matrix, where w_{ij} represents the importance of term i in biographic entry j

Dictionary	Biography 1	Biography 2	.	Biography n
University	w_{11}	w_{12}	.	w_{1n}
Marriage	w_{21}	w_{22}	.	w_{2n}
Novel	w_{31}	w_{32}	.	w_{3n}
.	.	.	.	.
Writer	w_{m1}	w_{m2}	.	w_{mn}

from the "Escritoras de España" category. In contrast, the first 416 pages of category "Escritores de España del siglo XX" have been retrieved for male biographies. Hence, these biographies are also divided into a female and a male set.

Table 1 summarises some general statistics from a preliminary exploratory analysis of the English and Spanish datasets described above.

We pre-process each biography using the *tm* text-mining R package Feinerer et al. (2008). Stop words (common words that usually do not add helpful information for the analysis) Eisenstein (2019), numbers, punctuation marks, and white spaces are removed from the biographies. Later, the text in every biography is converted to lowercase, and the corresponding term-biography matrix was created, as presented in Table 2. Thus, each biographical set can be represented as an $m \times n$ matrix, where m is the number of unique terms in the dictionary and n is the number of biographies in the training set. Each element w_{ij} of the term-biography matrix represents the importance or weight of the term i in the biography j. To obtain the value of w_{ij}, we use the TF-IDF measure Aizawa (2003), calculated as (17)

$$w_{ij} = tf_{ij} \times \log\left(\frac{n}{df_i}\right), \tag{17}$$

where tf_{ij} denotes the number of occurrences of the term i in the biography j; n is the total number of biographies in the training set; and, finally, df_i represents the number of biographies in which the term i appears.

We must remark that only training biographies contribute to the construction of the dictionary of terms that represents all biographies. That is, when the classifier checks new biographies in the testing set, any terms not found in biographies from the training set are simply ignored to calculate the distance. As well, synonyms, abbreviations, or alternative forms for a given term have not been considered in our study.

6 Results

This section presents the results of experiments comparing different strategies for text gender identification. Experiments for each classification algorithm involve ten trials of randomly selected train-test splits. To support the results, Wilcoxon statistical hypothesis tests Hollander et al. (2013) were performed, considering statistical significance for p values lower than 0.05.

Table 3 Average (avg.) and standard deviation (s.d.) of iteration count in training stages

Language	Method	Stemming			No stemming		
		avg.	s.d.	*p*-value	avg.	s.d.	*p*-value
Spanish	SVM	357.500	10.445	0.1	395.100	10.105	0.00195
	GSVM	294.100	89.709		293.000	49.844	
English	SVM	612.000	16.855	0.00195	614.600	7.394	0.00589
	GSVM	359.000	10.220		349.500	69.100	

6.1 GSVM versus standard SVM

The following sections summarise results from the evaluation experiments comparing GSVM with standard SVM Joachims (1998) in the context of text gender identification.

6.1.1 Iteration count comparison: GSVM versus standard SVM

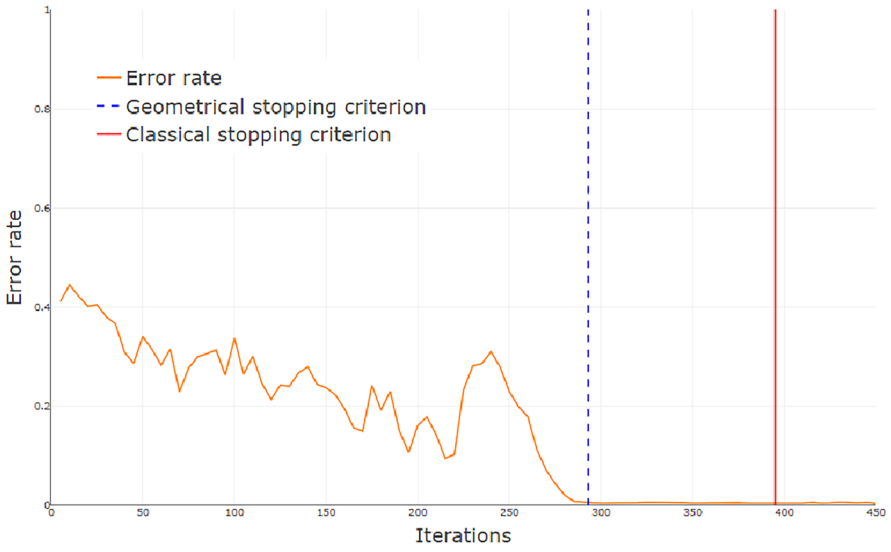
Table 3 compares the iteration count for GSVM and a typical SVM using the classical stopping criterion. It is clear that, on average, for the Spanish dataset, GSVM converges to approximately 18% (with stemming) and 25% (with no stemming) fewer iterations than the standard SVM approach. In turn, for the English dataset, GSVM reduces the number of iterations by approximately 42% (with stemming) and 44% (without stemming). The large values in the standard deviation for GSVM are a consequence of both approaches reaching the maximum number of iterations in some experiments, that is, the number of iterations of the standard SVM approach.

The Wilcoxon test for the English dataset shows that the improvement of GSVM over SVM is statistically significant, both when stemming is applied (p-value = 0.00195) and when it is not (p value = 0.00589). In the case of biographies in Spanish, the improvement is statistically significant when stemming is not included (p value = 0.00195), but not significant when stemming is used (p value = 0.1).

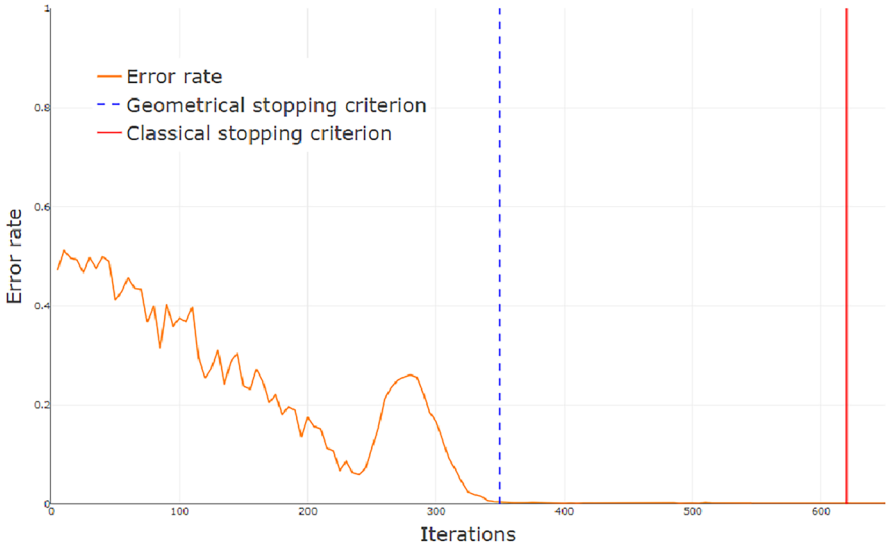
Figures 3a, 3b, 4a and 4b graphically show, for each experiment, at which iteration count GSVM and standard SVM stop their training execution, respectively. The GSVM stopping criterion detects the instant when the error rate stabilises (dotted line) in all cases. In contrast, the classical SVM stopping criterion requires a higher number of iterations to stop without providing any error rate improvement.

At this point, it is remarkable that the GSVM stopping criterion performance for biographies in English is similar, disregarding whether stemming pre-processing is applied or not (see Figs. 3b and 4b). Another noticeable finding is that the error rate for biographies in Spanish (see Figs. 3a and 4a) exhibits an oscillating pattern when the stemming pre-processing is applied (Fig. 4a). This behaviour may be due to a significant information loss introduced by the stemming operation in this language.

As a final comment, it is worth mentioning that the evaluation metrics are similar, in all cases, for the “female” and “male” output labels in both languages.

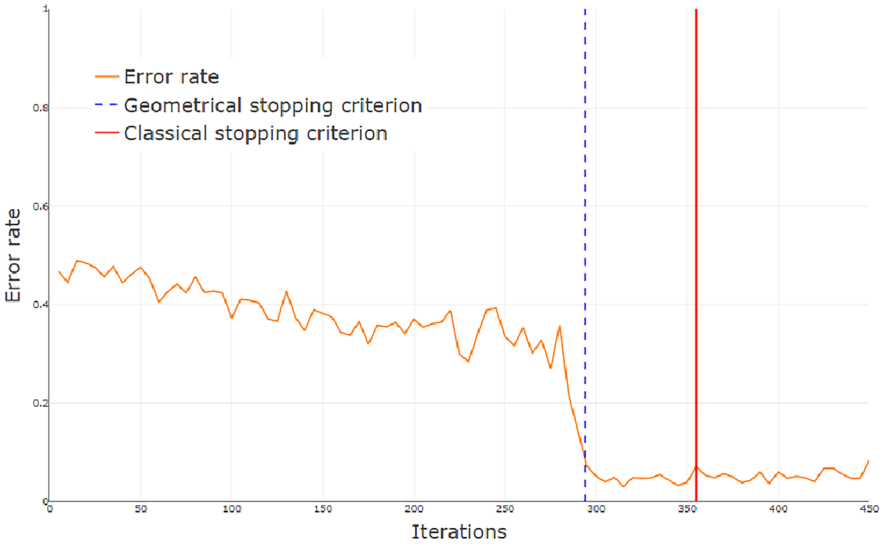


(a) Spanish language

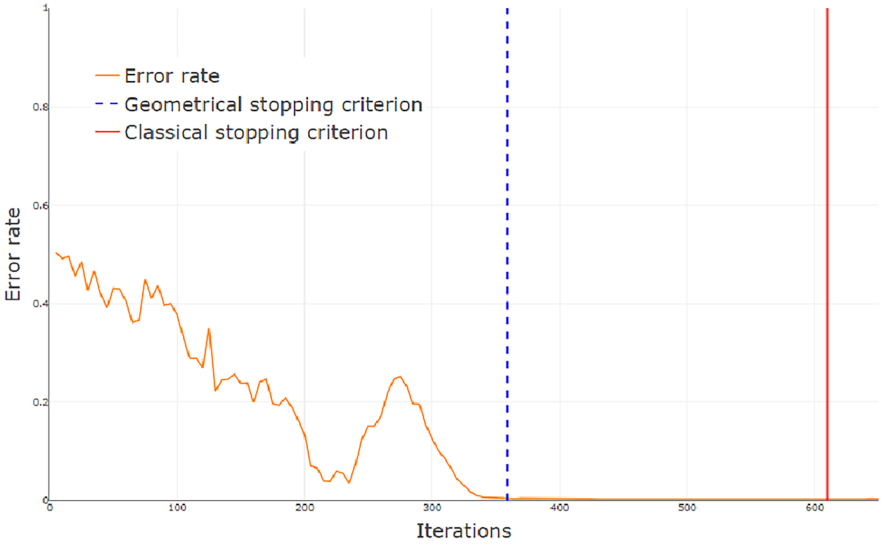


(b) English language

Fig. 3 **No stemming**. Comparison of the stopping criteria when the stemming pre-processing is not used



(a) Spanish language



(b) English language

Fig. 4 **Stemming**. Comparison of the stopping criteria when the stemming pre-processing is applied

Table 4 English. SVM and GSVM performance metrics for biographies in English classification, with and without stemming

Stem.	Gender	Method	Precision		Recall		F_1 score	
			avg. (s.d.)	p-value	avg. (s.d.)	p-value	avg. (s.d.)	p-value
No	Female	SVM	0.919 (0.038)	0.812	0.873 (0.011)	0.188	0.895 (0.021)	0.438
		GSVM	0.908 (0.039)		0.859 (0.093)		0.879 (0.049)	
	Male	SVM	0.865 (0.018)	0.63	0.917 (0.031)	1	0.889 (0.014)	1
		GSVM	0.823 (0.205)		0.909 (0.036)		0.843 (0.159)	
Yes	Female	SVM	0.911 (0.044)	0.812	0.879 (0.018)	0.0625	0.894 (0.027)	0.312
		GSVM	0.880 (0.049)		0.898 (0.036)		0.888 (0.031)	
	Male	SVM	0.873 (0.018)	0.125	0.910 (0.036)	1	0.891 (0.017)	0.438
		GSVM	0.899 (0.036)		0.884 (0.035)		0.891 (0.020)	

Table 5 Spanish. SVM and GSVM performance metrics for biographies in Spanish classification, with and without stemming

Stem.	Gender	Method	Precision		Recall		F_1 score	
			avg. (s.d.)	p-value	avg. (s.d.)	p-value	avg. (s.d.)	p-value
No	Female	SVM	0.929 (0.028)	1	0.918 (0.019)	0.0625	0.923 (0.014)	0.625
		GSVM	0.909 (0.065)		0.907 (0.093)		0.903 (0.042)	
	Male	SVM	0.913 (0.034)	0.188	0.929 (0.020)	1	0.920 (0.018)	0.188
		GSVM	0.889 (0.140)		0.913 (0.057)		0.891 (0.079)	
Yes	Female	SVM	0.751 (0.092)	0.0625	0.810 (0.072)	0.0625	0.773 (0.040)	0.625
		GSVM	0.770 (0.177)		0.781 (0.139)		0.747 (0.079)	
	Male	SVM	0.809 (0.115)	0.0625	0.768 (0.054)	0.125	0.780 (0.054)	0.0625
		GSVM	0.724 (0.251)		0.790 (0.105)		0.713 (0.155)	

6.1.2 Accuracy comparison: GSVM versus standard SVM

To evaluate the performance comparison between standard SVM and GSVM we follow the well-known metrics of *precision*, *recall* and F_1 *score* Baeza-Yates and Ribeiro-Neto (2011), Sokolova and Lapalme (2009), Olson and Delen (2008). Table 4 shows a comparison of SVM and GSVM for biographies in English, with and without stemming. For each method and metric, this table shows the average (avg.) and standard deviation (s.d.). In addition, the p-value of the Wilcoxon test is shown for each comparison. The Wilcoxon test results indicate that there are no statistically significant differences for any of the metrics considered, as shown in the p-value column, where all values are greater than 0.05. Results in Tables 4 and 5 empirically show that for text gender identification problems GSVM inherits the properties of SVM, providing similar accuracy results.

Table 6 Performance metrics' comparison of stemming versus no stemming when the GSVM method is applied

Language	Gender	Stem	Precision		Recall		F_1 score	
			avg	p-value	avg	p-value	avg	p-value
Spanish	Female	No	0.909	0.0097	0.907	0.00586	0.903	0.00195
		Yes	0.770		0.781		0.747	
	Male	No	0.889	0.0137	0.913	0.00195	0.891	0.00195
		Yes	0.724		0.790		0.713	
English	Female	No	0.908	0.343	0.859	0.124	0.879	1
		Yes	0.880		0.898		0.888	
	Male	No	0.823	0.108	0.909	0.286	0.843	0.636
		Yes	0.899		0.884		0.891	

Table 5 shows a comparison of SVM and GSVM for biographies in Spanish, with and without stemming. Again, for each method and metric, this table shows the average (avg.), the standard deviation (s.d.), and the corresponding p-values of the Wilcoxon test. Similarly to biographies in English, the Wilcoxon test results indicate that there are no statistically significant differences for any of the metrics (p values are greater than 0.05). As a consequence, the drop in the number of iterations achieved by GSVM does not affect the classification accuracy.

Table 6 shows for GSVM a comparison of stemming versus no stemming, along with the corresponding p-values. For biographies in Spanish, results show significant degradation when stemming is included, with p values 0.0097, 0.00586 and 0.00195, for precision, recall, and F_1 score, respectively, on Female gender; and p-values 0.0137, 0.00195 and 0.00195, for precision, recall, and F_1-score, respectively, on Male gender. In contrast, for biographies in English, p values for all metrics exceed the 0.05 significance level, indicating the absence of a significant effect.

6.2 GSVM versus other algorithms

Next, we compare the GSVM performance and training time against the performance and training time obtained for other well-known machine learning techniques, namely Random Forests (RF) Breiman (2001) and Boosting Schapire (1990), implemented in standard software libraries. In these tests, the algorithms hyperparameters are fixed to their default values. For every algorithm and language, we compare two alternative text-mining workflows for each metric: i) including a stemming step, like in most conventional text-mining applications; and ii) avoiding stemming, retaining any affixes that provide extra gender information in some languages. We use time as a comparison metric, given that there is no possibility of fairly comparing iteration counts, as a different iteration approach drives the implementation of each method.

Table 7 Average (avg.) and standard deviation (s.d.) for training time

Language	Method	Stemming		No stemming	
		avg. (s.d.)	p-value	avg. (s.d.)	p-value
Spanish	GSVM	3.096 (0.460)		2.889 (0.150)	
	Boosting	31.110 (1.001)	0.002	143.370 (289.894)	0.002
	RF	132.718 (46.366)	0.002	234.024 (71.339)	0.002
English	GSVM	3.799 (0.074)		3.883 (0.525)	
	Boosting	59.978 (6.650)	0.002	88.976 (18.596)	0.002
	RF	224.732 (17.882)	0.002	301.715 (6.568)	0.002

Table 8 Accuracy metrics for methods on biographies in Spanish

Method	Gender	Stem	Precision		Recall		F_1 score	
			avg. (s.d.)	p-value	avg. (s.d.)	p-value	avg. (s.d.)	p-value
GSVM	Female	No	0.909 (0.065)	0.00977	0.907 (0.093)	0.00586	0.903 (0.042)	0.00195
		Yes	0.770 (0.177)		0.781 (0.139)		0.747 (0.079)	
	Male	No	0.889 (0.140)	0.0137	0.913 (0.057)	0.00195	0.891 (0.079)	0.00195
		Yes	0.724 (0.251)		0.790 (0.105)		0.713 (0.155)	
Boosting	Female	No	0.944 (0.018)	0.00195	0.938 (0.017)	0.00195	0.941 (0.012)	0.00195
		Yes	0.769 (0.044)		0.694 (0.037)		0.729 (0.022)	
	Male	No	0.936 (0.020)	0.00195	0.943 (0.015)	0.00195	0.939 (0.011)	0.00195
		Yes	0.655 (0.038)		0.736 (0.053)		0.691 (0.027)	
RF	Female	No	0.957 (0.016)	0.00195	0.970 (0.022)	0.00195	0.963 (0.011)	0.00195
		Yes	0.816 (0.046)		0.790 (0.065)		0.800 (0.037)	
	Male	No	0.971 (0.021)	0.00195	0.956 (0.020)	0.00195	0.963 (0.012)	0.00195
		Yes	0.779 (0.066)		0.806 (0.048)		0.789 (0.036)	

6.2.1 Training time comparison: GSVM versus other algorithms

Table 7 summarises the execution time (in seconds) for all versions of the algorithms considered in this study. Using stemming for the Spanish language, GSVM is, on average, up to 10.05 times faster than the Boosting method and 42.87 times faster than RF, whereas, for the English language, GSVM is, on average, up to 15.79 times faster than Boosting and 59.15 times faster than RF. When stemming is not applied, for the Spanish language, GSVM is, on average, up to 49.63 times faster than Boosting and 81.00 times faster than RF. As for English, GSVM is, on average, up to 22.91 times faster than Boosting and 77.70 times faster than RF. In summary, based on the conducted Wilcoxon test, the p-value column indicates a statistically significant difference between the methods, showing that the GSVM training time is consistently lower than in the other methods.

6.2.2 Accuracy comparison: GSVM versus other algorithms

Since stemming does not appear to be significant when processing biographies in English, for the sake of space, we will focus the remainder of our analysis on biographies in Spanish. Table 8 shows a comparison of stemming versus no stemming, and the corresponding p values, for GSVM, Boosting, and Random Forest. GSVM obtains very promising results, especially when no stemming is used, with precision, recall, and F_1 score over 90% in most cases, although Boosting and Random Forest mostly improve GSVM results. This is due to the fact that both techniques belong to the so-called ensemble methods (Sagi and Rokach 2018). By training multiple models and integrating their predictions, these methods enhance the predictive performance of a single model. We must remark that, since GSVM behaves like a standard SVM in terms of accuracy, these results should be considered from the point of view of the importance of stemming versus no stemming in languages such as Spanish. This pre-processing technique also affects the results that would be obtained using other types of classifiers.

In addition, the p values in Table 8 show a significant degradation in performance for all algorithms when stemming is applied. These results consistently show that for the Spanish language, without stemming, precision, recall, and F_1-score are over 90% in most cases, whereas using stemming a degradation rounding of 15% occurs.

7 Discussion

In this section, we consider some explanations for detected differences in gender identification performance and ponder over possible limitations in our work.

7.1 Classification algorithms' comparison

Results suggest that controlling for stemming application in textual data preparation, performance metrics are pretty similar, regardless of the machine learning algorithm selected for classification. Nevertheless, there exist differences depending on the language used.

As a result, we can conclude that machine learning classification algorithms for gender identification can be directly affected by suffixes present in the datasets used to train the algorithms. Our results confirm that stemming pre-processing, a method that apparently simplifies the classification problem, may induce gender confusion to solve the problem at hand.

7.2 Impact of stemming elimination

One of the primary goals of the proposed framework is to explore if retaining morphological elements containing gender information in some languages, like Spanish, can

provide superior classification performance for gender identification in text. According to the results presented in Tables 6 and 8, there is quite a noticeable increment in classification performance in Spanish for the algorithms considered in our experiments when the stemming step is eliminated from the data preparation pipeline.

In contrast, results for the English language indicate minimal variation in the F_1 score metric between applying stemming or not in the data preparation process for any classification algorithm. This language has similar outcomes regarding precision and recall performance metrics.

Globally, these results underpin our initial consideration of avoiding stemming can positively impact gender identification in texts when content is available in languages providing additional morphological elements that bring in gender information, like Spanish. Moreover, they also, confirm indications from prior research about the importance of carefully selecting the most appropriate combination of data preparation tasks, especially working with languages different from English Uysal and Gunal (2014).

7.3 Limitations

The novel GSVM method introducing an early stopping criterion for the training phase works well for the specific problem of text gender identification. Therefore, further experiments should be conducted to assess the performance of GSVM in other classification tasks before recommending GSVM as a general SVM methodological innovation in machine learning.

As for the proposed framework for text gender identification, an important limitation of our experimental setting is that we circumscribe the target gender variable to a binary choice. Other authors have already raised this issue Hamidi et al. (2018), Keyes (2018). In fact, Keyes (2018) found that 55 out of 58 studies for automated gender identification with machine learning assumed a binary gender output. In this regard, one possible path for future research could be replacing our binary classification algorithms with alternative machine learning models that support continuous and/or multivariate targets. In such contexts, it would be interesting to evaluate whether gender suffixes still provide some advantages for gender identification.

Another limitation of our approach is that input text may not be available in a strongly inflected language, different from English. However, in that case, the text could be automatically translated from English to Spanish or other languages with similar properties through automated services. This translated version of the input text can provide extra gender information to improve classification performance. In the same way, further research is needed to confirm that these initial results comparing English and Spanish are replicable with input data in other alternative languages.

8 Conclusions

In this article, we present a new stopping criterion for support vector optimisation algorithms based on the geometrical properties of vector representation of text content to solve the problem of automated gender identification. For this particular

problem, we show that the proposed algorithm requires less time for training compared to standard SVM algorithms. Regarding the framework, rather than following conventional normalisation procedures in text mining that eliminate gender affixes, such as stemming, we retain those morphological elements found in strongly inflected languages to improve the performance of gender identification methods. We assess the effectiveness of this new approach in terms of training times by comparing different machine learning algorithms for classification, using a dataset of biographical entries from Wikipedia in English and Spanish.

Our results suggest that avoiding stemming positively influences gender tagging in text for languages that include additional elements incorporating gender information, like Spanish. This procedural change does not impact more gender-neutral languages like English.

Acknowledgements This work has been partially supported by grant TED2021-131295B-C33, funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR; and by grant XMIDAS, ref. PID2021-122640OB-I00, funded by the Spanish Ministry of Science and Innovation.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Data availability The English dataset is publicly available at https://doi.org/10.6084/m9.figshare.13551467. The Spanish dataset is available at https://doi.org/10.6084/m9.figshare.13551437.

References

Aggarwal CC (2018) Machine Learning For Text. Springer, Cham, Switzerland. https://doi.org/10.1007/978-3-319-73531-3

Breiman L (2001) Random Forests. Mach Learn 45(1):5–32. https://doi.org/10.1023/A:1010933404324

Adler BT, de Alfaro L, Mola-Velasco SM, Rosso P, West AG (2011) Wikipedia Vandalism Detection: Combining Natural Language, Metadata and Reputation Features. In: Gelbukh, A.F. (ed.) Computational Linguistics and Intelligent Text Processing - 12th International Conference, CICLing 2011, Tokyo, Japan, February 20-26, 2011. Proceedings, Part II. Lecture Notes in Computer Science, vol. 6609, pp. 277–288. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-19437-5_23

Aizawa A (2003) An information-theoretic perspective of tf-idf measures. Inform Process Manag 39(1):45–65. https://doi.org/10.1016/S0306-4573(02)00021-3

Amado A, Cortez P, Rita P, Moro S (2018) Research trends on big data in marketing: a text mining and topic modeling based literature analysis. Euro Res Manag Business Econ 24(1):1–7

Baeza-Yates R, Ribeiro-Neto B (2011) Modern information retrieval: the concepts and technology behind search, 2nd edn. Pearson Education Ltd., Harlow, England

Berry MW, Kogan J (eds) (2010) Text Mining: Applications and Theory. Wiley InterScience. John Wiley & Sons, Chichester, West Sussex, UK

Chen P-H, Fan R-E, Lin C-J (2006) A study on SMO-type decomposition methods for support vector machines. IEEE Trans Neural Netw 17(4):893–908

Cho H-C, Okazaki N, Miwa M, Tsujii J (2013) Named entity recognition with multiple segment representations. Inform Process Manage 49(4):954–965. https://doi.org/10.1016/j.ipm.2013.03.002

Corney M, de Vel OY, Anderson A, Mohay GM (2002) Gender-Preferential Text Mining of E-mail Discourse. In: 18th Annual Computer Security Applications Conference (ACSAC 2002), 9-13 December 2002, Las Vegas, NV, USA, pp. 282–289. IEEE Computer Society, Piscataway, NJ, USA. https://doi.org/10.1109/CSAC.2002.1176299

Das M, Hecht B, Gergle D (2019) The Gendered Geography of Contributions to OpenStreetMap: Complexities in Self-Focus Bias. In: Brewster, S.A., Fitzpatrick, G., Cox, A.L., Kostakos, V. (eds.) Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 4-9, 2019, pp. 1–14. ACM, New York, NY, USA. https://doi.org/10.1145/3290605.3300793

Das S, Paik JH (2021) Context-sensitive gender inference of named entities in text. Inform Process Manag 58(1):102423. https://doi.org/10.1016/j.ipm.2020.102423

Eisenstein J (2019) Introduction to Natural Language Processing. Adaptive Computation and Machine Learning series. MIT Press, Cambridge, MA, USA

Fatima M, Hasan K, Anwar S, Nawab RMA (2017) Multilingual author profiling on Facebook. Inform Process Manage 53(4):886–904. https://doi.org/10.1016/j.ipm.2017.03.005

Feldman R, Sanger J (2006) The Text Mining Handbook - Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, New York, NY, USA. https://doi.org/10.1017/CBO9780511546914

Feinerer I, Hornik K, Meyer D (2008) Text Mining Infrastructure in R. Journal of Statistical Software, Articles 25(5), 1–54. https://doi.org/10.18637/jss.v025.i05

Feng M, Li S (2018) An approximate strong kkt condition for multiobjective optimization. Top 26(3):489–509. https://doi.org/10.1007/s11750-018-0491-6

Foong E, Vincent N, Hecht B, Gerber EM (2018) Women (Still) Ask For Less: Gender Differences in Hourly Rate in an Online Labor Marketplace. Proc. ACM Hum.-Comput. Interact. 2(CSCW), 53–15321. https://doi.org/10.1145/3274322

Fourkioti O, Symeonidis S, Arampatzis A (2019) Language models and fusion for authorship attribution. Inform Process Manag 56(6):102061. https://doi.org/10.1016/j.ipm.2019.102061

Geiger RS, Ribes D (2010) The work of sustaining order in Wikipedia: the banning of a vandal. In: Inkpen, K., Gutwin, C., Tang, J.C. (eds.) Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work, CSCW 2010, Savannah, Georgia, USA, February 6-10, 2010, pp. 117–126. ACM, New York, NY, USA. https://doi.org/10.1145/1718918.1718941

Gomez J, Alfaro C, Ortega F, Moguerza JM, Algar MJ, Moreno R (2021). Biographies of literature writers written in English language. https://doi.org/10.6084/m9.figshare.13551467.v4. url figshare.com/articles/dataset/Biographies_of_literature_writers/13551467

Gomez J, Alfaro C, Ortega F, Moguerza JM, Algar MJ, Moreno R (2021). Biographies of literature writers written in Spanish language. https://doi.org/10.6084/m9.figshare.13551437.v5. url figshare.com/articles/dataset/biographies_RData/13551437

Hamidi F, Scheuerman MK, Branham SM (2018) Gender Recognition or Gender Reductionism? The Social Implications of Embedded Gender Recognition Systems. In: Mandryk, R.L., Hancock, M., Perry, M., Cox, A.L. (eds.) Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, Montreal, QC, Canada, April 21-26, 2018, pp. 1–13. ACM, New York, NY, USA. https://doi.org/10.1145/3173574.3173582

Hedlund T, Pirkola A, Järvelin K (2001) Aspects of Swedish morphology and semantics from the perspective of mono- and cross-language information retrieval. Inform Process Manag 37(1):147–161. https://doi.org/10.1016/S0306-4573(00)00024-8

Hollander M, Wolfe DA, Chicken E (2013) Nonparametric Statistical Methods. John Wiley & Sons, Bobooken, New Jersey

Huang F, Li C, Lin L (2014) Identifying Gender of Microblog Users Based on Message Mining. In: Li, F., Li, G., Hwang, S., Yao, B., Zhang, Z. (eds.) Web-Age Information Management - 15th International Conference, WAIM 2014, Macau, China, June 16-18, 2014. Proceedings. Lecture Notes in Computer Science, vol. 8485, pp. 488–493. Springer, Cham, Switzerland. https://doi.org/10.1007/978-3-319-08010-9_54

Jansen BJ, Moore K, Carman S (2013) Evaluating the performance of demographic targeting using gender in sponsored search. Inform Process Manag 49(1):286–302. https://doi.org/10.1016/j.ipm.2012.06.001

Joachims T (1998) Text categorization with Support Vector Machines: Learning with many relevant features. In: Nédellec, C., Rouveirol, C. (eds.) 10th European Conference on Machine Learning, ECML-98, Chemnitz, Germany, April 21-23, 1998. Proceedings. Lecture Notes in Computer Science, vol. 1398, pp. 137–142. Springer, Berlin, Heidelberg. https://doi.org/10.1007/BFb0026683

Joachims T (1999) Making Large-Scale Support Vector Machine Learning Practical. In: Schölkopf, B., Burges, C.J.C., Smola, A.J. (eds.) Advances in Kernel Methods: Support Vector Learning, pp. 169–184. MIT Press, Cambridge, MA, USA. Chap. 11

Joachims T (2002) Learning to Classify Text Using Support Vector Machines. The Springer International Series in Engineering and Computer Science, vol. 668. Springer, New York, NY, USA. https://doi.org/10.1007/978-1-4615-0907-3

Jurafsky D, Martin JH (2009) Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, And Speech Recognition, 2nd edn. Prentice Hall Series in Artificial Intelligence. Prentice Hall, Pearson Education International, London, UK. https://www.worldcat.org/oclc/315913020

Keyes O, Tilbert B (2017) WikipediR: A MediaWiki API Wrapper. R package version 1.5.0. https://CRAN.R-project.org/package=WikipediR

Keyes O (2018) The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. Proc. ACM Hum.-Comput. Interact. 2(CSCW), 88–18822. https://doi.org/10.1145/3274357

Kocher M, Savoy J (2017) Distance measures in author profiling. Inform Process Manag 53(5):1103–1119. https://doi.org/10.1016/j.ipm.2017.04.004

Kretschmer H, Aguillo IF (2005) New indicators for gender studies in web networks. Information Processing & Management 41(6):1481–1494. https://doi.org/10.1016/j.ipm.2005.03.009. Special Issue on Infometrics

Krüger S, Hermann B (2019) Can an Online Service Predict Gender? On the State-of-the-Art in Gender Identification from Texts. In: Crnkovic, I., Silveira, K.K., Sprenkle, S. (eds.) Proceedings of the 2nd International Workshop on Gender Equality in Software Engineering, GE@ICSE 2019, Montreal, QC, Canada, May 27, 2019, pp. 13–16. IEEE Press, Piscataway, NJ, USA. https://doi.org/10.1109/GE.2019.00012

Kucukyilmaz T, Cambazoglu BB, Aykanat C, Can F (2006) Chat Mining for Gender Prediction. In: Yakhno, T.M., Neuhold, E.J. (eds.) Advances in Information Systems, 4th International Conference, ADVIS 2006, Izmir, Turkey, October 18-20, 2006, Proceedings. Lecture Notes in Computer Science, vol. 4243, pp. 274–283. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11890393_29

Lau K-N, Lee K-H, Ho Y (2005) Text Mining for the Hotel Industry. Cornell Hotel Restaurant Administration Q 46(3):344–362. https://doi.org/10.1177/0010880405275966

Lin B, Serebrenik A (2016) Recognizing gender of Stack Overflow users. In: Kim, M., Robbes, R., Bird, C. (eds.) Proceedings of the 13th International Conference on Mining Software Repositories, MSR 2016, Austin, TX, USA, May 14-22, 2016, pp. 425–429. ACM, New York, NY, USA. https://doi.org/10.1145/2901739.2901777

López-Santillán R, Montes-Y-Gómez M, González-Gurrola LC, Ramírez-Alonso G, Prieto-Ordaz O (2020) Richer document embeddings for author profiling tasks based on a heuristic search. Inform Processing Manag 57(4):102227. https://doi.org/10.1016/j.ipm.2020.102227

Markov I, Gómez-Adorno H, Sidorov G, Gelbukh A (2017) The Winning Approach to Cross-Genre Gender Identification in Russian at RUSProfiling 2017. In: Majumder, P., Mitra, M., Mehta, P., Sankhavara, J. (eds.) Working Notes of FIRE 2017 - Forum for Information Retrieval Evaluation, Bangalore, India, December 8-10, 2017. CEUR Workshop Proceedings, vol. 2036, pp. 20–24. CEUR-WS.org, Aachen, Germany. http://ceur-ws.org/Vol-2036/T1-5.pdf

Moguerza JM, Muñoz A et al (2006) Support vector machines with applications. Stat Sci 21(3):322–336. https://doi.org/10.1214/088342306000000493

Mukherjee S, Bala PK (2017) Gender classification of microblog text based on authorial style. Inform Syst e-Business Manag 15(1):117–138. https://doi.org/10.1007/s10257-016-0312-0

Olson DL, Delen D (2008) Advanced data mining techniques. Springer, Berlin, Heidelberg

Platt J (1998) Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. Technical Report MSR-TR-98-14, Microsoft. https://www.microsoft.com/en-us/research/publication/sequential-minimal-optimization-a-fast-algorithm-for-training-support-vector-machines/

Rangel F, Rosso P (2016) On the impact of emotions on author profiling. Inform Process Manag 52(1):73–92. https://doi.org/10.1016/j.ipm.2015.06.003
R Core Team (2022) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. R Foundation for Statistical Computing. https://www.R-project.org/
Santamaría L, Mihaljević H (2018) Comparison and benchmark of name-to-gender inference services. PeerJ Comput Sci 4:156. https://doi.org/10.7717/peerj-cs.156
Sagi O, Rokach L (2018) Ensemble learning: a survey. Wiley interdisciplinary reviews: data mining and knowledge discovery 8(4):1249
Schapire RE (1990) The Strength of Weak Learnability. Mach Learn 5(2):197–227. https://doi.org/10.1007/BF00116037
Sokolova M, Lapalme G (2009) A systematic analysis of performance measures for classification tasks. Inform Process Manag 45(4):427–437
Sproat R, Black AW, Chen S, Kumar S, Ostendorf M, Richards C (2001) Normalization of non-standard words. Comput Speech Lang 15(3):287–333. https://doi.org/10.1006/csla.2001.0169
Srivastava A, Sahami M (2009) (eds.): Text Mining: Classification, Clustering, and Applications, 1st edn. Chapman & Hall/CRC, New York, NY, USA. https://doi.org/10.1201/9781420059458
Tikhonov AN, Arsenin VY (1977) Solutions of Ill-Posed Problems. Scripta Series in Mathematics. Halsted Press, John Wiley & Sons, New York, NY, USA
Terrell J, Kofink A, Middleton J, Rainear C, Murphy-Hill E, Parnin C, Stallings J (2017) Gender differences and bias in open source: pull request acceptance of women versus men. PeerJ Comput Sci 3:111. https://doi.org/10.7717/peerj-cs.111
Uysal AK, Gunal S (2014) The impact of preprocessing on text classification. Inform Process Manag 50(1):104–112. https://doi.org/10.1016/j.ipm.2013.08.006
Vasilescu B, Capiluppi A, Serebrenik A (2014) Gender, representation and online participation: a quantitative study. Interacting Comput 26(5):488–511. https://doi.org/10.1093/iwc/iwt047
Vrandečić D, Krötzsch M (2014) Wikidata: a free collaborative knowledgebase. Commun ACM 57(10):78–85. https://doi.org/10.1145/2629489
Wais K (2016) Gender Prediction Methods Based on First Names with genderizeR. The R Journal 8(1), 17–37. https://doi.org/10.32614/RJ-2016-002
Witten IH, Frank E, Hall MA, Pal CJ (2017) Data Mining: Practical Machine Learning Tools and Techniques, 4th edn. Morgan Kaufmann, Elsevier, Cambridge, MA, USA. https://doi.org/10.1016/C2015-0-02071-8
Yan X, Yan L (2006) Gender Classification of Weblog Authors. In: Computational Approaches to Analyzing Weblogs, Papers from the 2006 AAAI Spring Symposium, Technical Report SS-06-03, Palo Alto, CA, USA, March 27-29, 2006, pp. 228–230. AAAI Press, Palo Alto, CA, USA. https://aaai.org/papers/0046-gender-classification-of-weblog-authors/

Authors and Affiliations

Javier Gomez[1] · Cesar Alfaro[1] · Felipe Ortega[1] · Javier M. Moguerza[1] · Maria Jesus Algar[1] · Raul Moreno[2]

✉ Javier Gomez
javier.gomez@urjc.es

✉ Raul Moreno
r.morenoi@alumnos.urjc.es

Cesar Alfaro
cesar.alfaro@urjc.es

Felipe Ortega
felipe.ortega@urjc.es

Javier M. Moguerza
javier.moguerza@urjc.es

Maria Jesus Algar
mariajesus.algar@urjc.es

[1] Research Centre for Intelligent Information Technologies (CETINIA-DSLAB), Rey Juan Carlos University, Calle Tulipán, Móstoles 28933, Madrid, Spain

[2] Doctorate Programme in Information Technologies and Communications, Rey Juan Carlos University, Calle Tulipán, Móstoles 28933, Madrid, Spain