applied sciences

MDPI

*Article*

# A White-Box Sociolinguistic Model for Gender Detection

**Damián Morales Sánchez [1,*], Antonio Moreno [2] and María Dolores Jiménez López [1]**

[1] Research Group on Mathematical Linguistics (GRLMC), Universitat Rovira i Virgili, 43002 Tarragona, Spain; mariadolores.jimenez@urv.cat
[2] Intelligent Technologies for Advanced Knowledge Acquisition (ITAKA), Universitat Rovira i Virgili, 43007 Tarragona, Spain; antonio.moreno@urv.cat
* Correspondence: damian.morales@estudiants.urv.cat

**Abstract:** Within the area of Natural Language Processing, we approached the Author Profiling task as a text classification problem. Based on the author's writing style, sociodemographic information, such as the author's gender, age, or native language can be predicted. The exponential growth of user-generated data and the development of Machine-Learning techniques have led to significant advances in automatic gender detection. Unfortunately, gender detection models often become black-boxes in terms of interpretability. In this paper, we propose a tree-based computational model for gender detection made up of 198 features. Unlike the previous works on gender detection, we organized the features from a linguistic perspective into six categories: orthographic, morphological, lexical, syntactic, digital, and pragmatics-discursive. We implemented a Decision-Tree classifier to evaluate the performance of all feature combinations, and the experiments revealed that, on average, the classification accuracy increased up to 3.25% with the addition of feature sets. The maximum classification accuracy was reached by a three-level model that combined lexical, syntactic, and digital features. We present the most relevant features for gender detection according to the trees generated by the classifier and contextualize the significance of the computational results with the linguistic patterns defined by previous research in relation to gender.

**Keywords:** gender detection; machine learning; author profiling; computational sociolinguistics



**Citation:** Morales Sánchez, D.; Moreno, A.; Jiménez López, M.D. A White-Box Sociolinguistic Model for Gender Detection. *Appl. Sci.* **2022**, *12*, 2676. https://doi.org/10.3390/app12052676

Academic Editor: Evgeny Nikulchev

Received: 31 December 2021
Accepted: 2 March 2022
Published: 4 March 2022



## 1. Introduction

In recent years, User-Generated Content (UGC) has increased due to the proliferation of web spaces that encourage user participation, such as blogs or social networks: "Gradually, a greater range of tools and platforms for the development and hosting of such content emerged, resulting in a further widening of participation in user-led content creation" [1].

Computational social sciences [2] consider UGC as a valuable source of information to understand social phenomena and reinforce their sociological explanations with quantitative analyses [3]. Within this research agenda, Author Profiling (AP) analyzes the authors' linguistic productions in order to trace identifying clues, such as age, gender, personality traits, or native language. Gender detection models developed by the AP field directly benefit other research areas, such as forensic linguistics, marketing, and sociolinguistics [4].

The World Wide Web provides an environment in which new types of criminal activities are perpetrated [5]. Specifically, online harassment has been closely related to anonymity, since users can join virtual communities, such as social networks by providing a false identity using nicknames or false pictures: "Anonymity has long been thought to encourage bad behavior, either by changing the salient norms, or through reducing the subjective need to adhere to norms by dampening the effect of internal mechanisms such as guilt and shame" [6].

From a forensic linguistic point of view, the linguistic choices made by a suspected author become textual evidence in legal investigation. In this vein, researchers have designed

computational models to automatically detect different types of online harassment, such as sexual harassment [7] or cyberbullying [8].

However, AP models have gradually moved away from the needs of forensic linguistics since they have lost explanatory power with the implementation of black-box algorithms. Forensic linguists ultimately need language as evidence, and therefore profiling models cannot incur opacity in exchange for accuracy: "Computational authorship profiling is not necessarily interested in understanding the inner (linguistic) mechanisms of the machine, as long as the accuracy rates are outperforming previous models" [9].

In addition, marketing companies are interested in developing recommender systems from the sociodemographic data of their customers in order to improve customer experience making personalized recommendations of products and services [10]. Recommender systems are based on Machine-Learning algorithms that learn from the textual data generated by the users through comments and reviews or from their browsing activities in shopping applications or websites. In these marketing computational applications, gender has been considered a useful social variable [11]. For example, in [12], Aljohani and Cristea designed a Deep-Learning model to detect the gender of MOOC learners in order to offer them personalized course content.

Finally, in [13] Nguyen et al. indicated the contribution of automatic gender detection on sociolinguistics, traditionally concerned with defining linguistic patterns correlated with social variables, such as age, gender, or social class. Sociolinguistic conclusions related to gender were drawn mainly from face-to-face informal interactions. Consequently, the introduction of Machine-Learning techniques for the analysis of large amounts of textual data may contribute to the definition of new sociolinguistic patterns.

This study contributes to sociolinguistics and to computational sociolinguistics in particular, as well as related research fields, since we propose a computational model for gender detection in Spanish based on decision trees. Beyond the classification task, our objective is twofold: on the one hand, to verify previous sociolinguistic conclusions about gender and, on the other hand, to define new linguistic patterns that may contribute to disciplines, including forensic linguistics, in profiling the gender of authors from textual data.

The style and linguistic characteristics of the messages are traces that the author leaves. They allow unmasking the identity of the person who hides behind the text even when the perpetrator poses as another identity. Our model is committed to interpretability, since the forensic linguist needs access to the linguistic traces that allow the identification of the suspect as evidence for a legal investigation.

We incorporated sociolinguistic features, such as tag questions or appellatives, along with other features previously explored in the field of author profiling, such as lexical richness. Unlike previous works, we handled a reduced set of features, specifically, less than 200 features, which covered a wide linguistic spectrum, from orthography to pragmatics. In fact, despite that our model reached 5.53% less accuracy than the model of Santosh et al. in [14], we obtained competitive results handling less than 1% of the number of features.

The paper is organized as follows. Section 2 reviews the state of the art regarding automatic gender detection. Section 3 describes our tree-based computational model for gender detection. In Section 4, we report and discuss the experimental results. The paper finishes with some concluding remarks and future research avenues presented in Section 5.

## 2. Automatic Gender Detection: From Discriminant Analysis to Deep Learning

Gender has been, by far, the sociodemographic characteristic that has received the most attention both in computational studies and in sociolinguistic research. Within the computational field, automatic gender detection began around the turn of the new millennium. Researchers implemented Artificial-Intelligence methods, such as Machine-Learning algorithms, to classify sets of authors according to their gender. Unlike sociolinguistics, which incorporated sociological theories about gender in its analyses, automatic gender detection has been approached as a binary classification task (female or male).

This simplistic notion of gender was questioned in [15]: "If we start with the assumption that 'female' and 'male' are the relevant categories, then our analyses are incapable of revealing violations of this assumption". In relation to automatic gender detection studies, four stages may be identified.

### 2.1. First Stage

Early automatic gender detection studies examined formal texts, such as the British National Corpus or the Enron e-mail corpus, or datasets collected from controlled settings. Research on automatic gender detection was constrained due to the unavailability of annotated data.

In [16], Thomson and Murachver conducted three experiments to analyze gender-preferential language style in electronic discourse. They computed the number of messages, the total word count, the sentence length as well as other lexical and morphological features, such as references to emotion, compliments, insults, apologies, intensive adverbs, subordinating conjunctions, and adjectives, among others. Although they detected few significant differences in the electronic messages written by the 35 participants, they correctly classified 91.4% of the authors using a discriminant analysis.

Singh [17] examined gender differences related to lexical richness in free and spontaneous recorded conversations produced by 30 participants. Specifically, he applied eight lexical richness measures, such as noun-rate and adjective-rate per 100 words, and type-token ratio. By performing a statistical discriminant analysis, Singh reached a classification accuracy of 90%.

In 2002, Corney et al. [18] analyzed an e-mail corpus sourced from a large academic organization. Initially, they collected 8820 e-mail documents written by 342 authors. They reported a 70.2% $F_1$ score with a Support Vector Machine classifier and 221 features, organized into seven categories: document-based, word-based, character-based, function words, structural, gender-preferential, and other features. The so-called gender-preferential features mainly consisted of the frequency of adjectives and adverbs, along with the frequency of the word *sorry* and apology-related words.

Koppel, Argamon and Shimoni [19] conducted a gender detection study on 566 documents from the British National Corpus labeled both for author gender and for genre (fiction and non-fiction). They proposed a learning method based on the Exponential Gradient algorithm to find a linear separator between female-authored and male-authored texts. They used 405 function words and the 500 most common ordered triples, the 100 most common ordered pairs and all the single parts-of-speech tags as features. With this method, they reached approximately 80% accuracy inferring the authors' gender and 98% on genre identification.

In 2005, Boulis and Ostendorf [20] performed a computational analysis on the Fisher corpus made up of 12,000 recorded telephone conversations. They extracted word unigrams and bigrams as features, and they tested various Machine-Learning algorithms, such as Naïve Bayes, Maximum Entropy, and Rocchio. The maximum accuracy of 92.5% was achieved by a Support Vector Machine classifier using about 300 K word bigrams.

The consolidation of the blogosphere with the expansion of platforms, including blogger.com or wordpress.org, represented a paradigm shift in automatic gender detection.

### 2.2. Second Stage

In a second stage, automatic gender detection moved from formal documents to informal texts due to the possibility of collecting large datasets from the blogosphere. Moreover, researchers considered, along with gender, other sociodemographic characteristics, such as age, origin, and personality traits.

Nowson and Oberlander [21] collected a personal weblog corpus constituted by 71 authors to predict the authors' gender and also to evaluate their openness attitude. Their features consisted of dictionary-based features (Linguistic Inquiry and Word Count and the

MRC psycholinguistic database) and 125 n-grams of parts-of-speech tags. With a Support Vector Machine classifier, they reached 92.5% classification accuracy for gender detection.

In [22], Schler et al. examined the effects of gender and age on a blog dataset sourced from blogger.com made up of 71,000 blogs. However, to prevent bias, they created a subdataset of 37,478 blogs guaranteeing an equal gender composition in each age group. Schler et al. computed the post length and extracted function words, parts-of-speech tags, blog words, and hyperlinks as features. They reported 80.1% classification accuracy using the algorithm Multi-Class Real Winnow.

Yan and Yan [23] collected 75,000 blog entries written by 3000 authors from Xanga. They extracted unigrams along with weblog-specific features, such as background color, word fonts and cases, and emoticons. They reported a 68% $F_1$ score on gender detection with a Naïve Bayes classifier.

In 2009, Goswami, Sarkar and Rustagi [24] predicted gender and age from a 20,000 blog corpus previously explored by [22], considering the frequency of 52 non-dictionary words and the length of sentences. With a Naïve Bayes classifier, they achieved 89.3% classification accuracy on gender detection.

In [25], Mukherjee and Liu collected their own blog corpus made up of 3100 blogs sourced from various blog hosting sites, such as technorati.com and blogger.com. They extracted stylistic features, word classes, and gender-preferential features [18], along with parts-of-speech sequence patterns. With the implementation of a Support Vector Machine classifier, they reached 88.56% classification accuracy.

Otterbacher [26] explored a movie review corpus composed of 31,300 reviews sourced from the Internet Movie Database (IMDb) website. He applied a Logistic Regression classifier to content-based and metadata features to reach a classification accuracy of 73.3%.

### 2.3. Third Stage

In a third stage, automatic gender detection studies were interested in microblogging. The previous stages demonstrated that it was possible to predict the gender of authors from formal and informal texts of a certain length; however, microblogging platforms, such as Twitter, set a new challenge: to extract identifying information with little textual input.

The first study on Twitter was conducted by Rao et al. [27]. They manually annotated 1000 Twitter users to infer four user attributes: gender, age, political orientation, and regional origin. They achieved a classification accuracy of 72.3% using a Support Vector Machine classifier with unigrams and bigrams tokens (1,256,558 features) and sociolinguistic features. The latter set of features mainly consisted of punctuation mark frequencies, a list of emoticons compiled from the Wikipedia, and some words lists, such as exasperation expressions or affection words.

In [28], Burger, Henderson, Kim and Zarrella created a large dataset of 184,000 Twitter users with 4.1 million tweets for training. Unlike the work presented by [27], they automatically annotated their dataset following the URLs included by the users in their profile description that linked to their blog sites. They experimented with a wide variety of classifiers, including Naïve Bayes, Support Vector Machines, and Balanced Winnow 2. They reported 75.5% classification accuracy with a Balanced Winno2 classifier using character n-grams with $n$ in the range (1, 5) and word unigrams and bigrams (15,572,522 features).

In 2012, Fink, Kipecky and Morawski [29] collected 78,853 Twitter users in order to infer authors' gender from the content of their tweets. They reached 80.6% classification accuracy using a Support Vector Machine classifier with word unigrams, Twitter hastags and LIWC categories as features (1,231,910 features).

Ciot, Sonderegger and Ruths [30] also experimented with a Support Vector Machine classifier on Twitter data. However, unlike previous work, they extracted tweets written in four languages: Japanese, Indonesian, French, and Turkish. They reached the highest accuracy of 87% on Turkish and the lowest accuracy of 63% on Japanese with the 20 top words uniggrams, bigrams, trigrams, and hashtags, along with other metadata features, such as the 20 top mentions or links.

Alrifai, Rebdawi and Ghneim [31], in 2017, implemented the Sequential Minimal Optimization algorithm on the PAN-AP-17 dataset. They reported 72.25% accuracy on gender detection using character n-grams with $n$ in range (2, 7).

### 2.4. Fourth Stage

The latest developments in Machine Learning have been incorporated in recent years in automatic gender detection research. More specifically, Deep Learning structures have been designed to perform gender detection from textual and visual data.

In [32], Manna, Pascucci and Monti analyzed 56 blogs collected from the SogniLucidi website using a Feed-Forward Neural Network. They reported 77.6% classification accuracy using word unigrams, bigrams, and trigrams as features.

Park and Woo [33] explored an AIDS-related bulletin board from HealthBoard.com and created a gender detection model based on the emotions expressed by the users in their comments and posts. They applied the NRC and BING dictionary to extract sentiment information. The results exhibited that Deep-Learning structures, and more specifically, Convolutional Neural Networks, outperformed traditional Machine-Learning algorithms. In fact, they reached 73.44% accuracy with Random Forest, whereas a Convolutional Neural Network structure yielded 91% accuracy.

In 2020, Safara et al. [34] explored the Enron e-mail corpus with an Artificial Neural Network and the Whale Optimization Algorithm to find the optimal weights and improve the accuracy of the neural network structure. They outperformed previous work on the Enron corpus achieving an accuracy of 98% with 48 linguistic features distributed into four categories: character-based features, word-based features, syntax-based features, and structure-based features.

Finally, Kowsari et al. [35] employed Deep Neural Networks and Convolutional Neural Networks on the PAN-AP-17 dataset collected from Twitter, reporting an accuracy of 86.33% using TF-IDF scores and GloVe word vectors.

Deep Learning methods have been implemented in other data modalities. To mention some examples, refs. [36,37] employed Convolutional Neural Network (CNN) for gender detection from facial images, and [38] used Multi-Scale Convolutional Neural Networks on audio data.

Table 1 presents an overview on some of the previous works on automatic gender detection. For each of the works reported, the table shows the dataset, features, and algorithm used as well as the accuracy reached.

**Table 1.** Previous work on automatic gender detection.

| Author | Year | Dataset | Features | Algorithm | Acc. |
|---|---|---|---|---|---|
| Singh | 2001 | Oral conversations | Lexical richness measures | Discriminant analysis | 90.0 |
| Boulis & Ostendorf | 2005 | Telephone conversations | Word unigrams and bigrams | SVM | 92.5 |
| Nowson & Oberlander | 2006 | Blogs | Dictionary-based and POS n-grams | SVM | 92.5 |
| Goswami, Sarkar & Rustagi | 2009 | Blogs | Non-dictionary words and sentence length | Naïve Bayes | 89.3 |
| Otterbacher | 2010 | Movie reviews | Lexical frequencies and POS tags | Logistic Regression | 73.3 |
| Rao et al. | 2010 | Twitter | Tokens unigrams and bigrams | SVM | 72.3 |
| Alrifai, Rebdawi & Ghneim | 2017 | Twitter | Character n-grams (2, 7) | Sequential Minimal Optimization | 72.25 |

**Table 1.** *Cont.*

| Author | Year | Dataset | Features | Algorithm | Acc. |
|---|---|---|---|---|---|
| Fink, Kipecky & Morawski | 2012 | Twitter | Word unigrams, hashtags and LIWC categories | Balanced Winno2 | 75.5 |
| Manna, Pascucci & Monti | 2019 | Blogs | Word unigrams, bigrams, and trigrams | Feed-Forward Neural Network | 77.6 |
| Park & Woo | 2019 | Web forum | NRC and BING dictionaries | CNN | 91 |
| Kowsari et al. | 2020 | Twitter | TF-IDF and GloVe | CNN | 86.33 |

## 3. A Sociolinguistic Model for Gender Detection

We implemented Machine-Learning techniques to explore social media posts and to design a computational model for gender detection. Our main objective was to define sociolinguistic patterns related to gender, and thus we prioritized the interpretability of the model over the classification accuracy.

### *3.1. PAN-AP-13 Dataset*

Our computational analysis was conducted on the dataset provided by the PAN organizing committee for the Author Profiling task at the Conference and Labs of the Evaluation Forum (CLEF) 2013 edition that took place in Valencia (Spain). PAN is conceived as "a series of scientific events and shared tasks on digital forensics and stylometry" (https://pan.webis.de/, accessed on 15 December 2021). Research teams participate in the shared tasks under the same computational conditions as they must deploy their models on the TIRA platform.

The PAN-AP-13 dataset consisted of social media posts written in English and Spanish, and labeled by gender (female and male) and age (10 s: 13–17 years; 20 s: 23–27 years; and 30 s: 33–47 years). We ran the experiments on the Spanish subdataset made up of 84,060 authors: 75,900 authors for training and 8160 for testing. We focused only on gender detection, and thus we omitted the information related to age. Table 2 presents the dataset distribution. We recommend [39] for more information about the data.

In the 2013 PAN AP task, 21 research teams participated. The classification accuracy on gender detection in the Spanish subdataset ranged from 47.84% to 64.73%. The maximum classification accuracy was reached by [14] using a Decision-Tree classifier trained with a combination of style-based features, such as punctuation marks frequencies, and content-based features, such as word unigrams and Latent Dirichlet Analysis-based topics.

**Table 2.** PAN-AP-13 Spanish dataset distribution.

| Language | Age Group | Gender | Number of Authors | |
|---|---|---|---|---|
| | | | Training | Test |
| Spanish | 10 s | M | 1250 | 144 |
| | | F | 1250 | 144 |
| | 20 s | M | 21,300 | 2304 |
| | | F | 21,300 | 2304 |
| | 30s | M | 15,400 | 1632 |
| | | F | 15,400 | 1632 |
| TOTAL | | | 75,900 | 8160 |

### 3.2. Features

For gender detection, researchers frequently organized the features according to the classification proposed in [4] by Argamon et al., who defined two basic types of features: style-based features and content-based features. Later, in [40] Neal et al. suggested a classification based on six categories: lexical, syntactic, semantic, structural, domain-specific, and additional features.

However, as our objective is to reveal linguistic patterns correlated to gender, we structured the features from a linguistic point of view in the following categories: orthographic, morphological, lexical, syntactic, and pragmatic-discursive. In addition, taking into account that the dataset contained digital elements, such as URLs or emoticons, we included a digital level. All features were extracted automatically with Python. In what follows, we introduce the different categories of the 198 features. Note that we indicate the number of features of each category in parentheses:

Orthographic features (29) mainly captured punctuation marks frequencies, such as single, double and angular quotation marks, commas, full-stops, colons, semi-colons, question marks, exclamation marks, parentheses, dashes, and ellipsis points. We also extracted sequences of punctuation marks, such as duplication of question and exclamation marks, repetition of question and exclamation marks, and combinations of question and exclamation marks. Finally, we computed frequencies of alphabetic characters, repetition of vowels and consonants, upper-case and lower-case characters, combination of upper- and lower-case characters, and numeric characters.

Morphological features (29) consisted of part-of-speech (PoS) tags. The word categories were extracted using petraTag (http://www.opentranslation.es/petratag/, accessed on 16 December 2021), an application developed by OpenTranslation and designed to tag corpora and obtain morphological information. Specifically, we counted the following PoS tags frequencies: nouns, adjectives, determiners (definite articles, demonstrative, possessive, indefinite, numeral, interrogative, and exclamative), pronouns (personal, demonstrative, possessive, indefinite, interrogative, exclamative, and relative), verbs (personal and non-personal verbs and verbal periphrasis), conjunctions (coordinating and subordinating conjunctions), prepositions, and interjections.

Lexical features (33) captured word-based information. We applied three dictionaries to extract information about the frequencies of slang, emotive and offensive words. More specifically, we used the Spanish Specific Lexicon of Social Networks and the Spanish Emotion Lexicon created by Sidorov (https://www.cic.ipn.mx/~sidorov/, accessed on 16 December 2021) [41] and a hand-crafted offensive word list. We also captured the use of mitigating lexical elements from the frequencies of modal and epistemic verbs, probability adjectives and adverbs, approximators, conditional tense verbs, subjunctive mood verb forms, and non-personal verbs, among other elements.

Syntactic features (29) were based on syntactic dependencies. We used the *Spacy* (https://spacy.io/, accessed on 16 December 2021) library to segment the messages into sentences and obtain the dependencies tags. We captured the following syntactic dependencies: nominal subjects, clausal subjects, direct objects, indirect objects, oblique nominal complements, nominal modifiers, adjectival modifiers, adverbial modifiers, numeric modifiers, determiners, case markers, appositional modifiers, clausal complements, open clausal complements, adverbial clause complements, and adjectival clause complements, along with syntactic relationships, such as coordination, juxtaposition, and subordination. In addition, we included other features, such as number of sentences, sentence length, word repetition by coordination, and the frequency of direct and indirect objects at the initial position.

Digital features (16) captured the frequencies of digital elements, such as URLs, embedded pictures, and emoticons. We computed the ratio between emoticons and words, and some emoticon frequencies related to basic emotions, such as sadness or joy.

Pragmatic-discursive features (62) captured pragmatics information, such as presuppositions and speech acts. More specifically, we extracted five types of explicit speech acts

(assertive, directive, commissive, expressive, and declaration) from the verbs of the sentences. For example, sentences formed by the verb *I promise* were captured as commissive speech acts, while sentences formed by *I order* were classified as directive speech acts. We created five lists of verbs based on the speech act theory.

With respect to presuppositions, we extracted existential (determiner phrases with definite interpretation, such as *the phone*, deictic terms and proper names), lexical (factive verbs, such as *regret*, verbs of judging, such as *criticizing*, change of state verbs and implicative verbs), and focal presuppositions (grammatical structures formed by a focus adverb, such as *even*). We also considered some features previously explored by sociolinguistics, such as tag questions, and politeness and apology expressions. Finally, we included some features in order to capture discursive information: discursive markers, type of sentences, and total number of words, line breaks, and tabulations.

### 3.3. Decision Trees for Gender Detection

Decision tree-based models have been applied for solving practical problems, such as medical diagnosis or marketing personalization. Unlike other Machine-Learning algorithms frequently employed in the AP area, such as Support Vector Machine, Naïve Bayes, and Deep Learning structures, Decision Trees are considered white-box models [42] because they generate interpretative and understandable models [43]: "For knowledge-based systems, decision trees have the advantage of being comprehensible by human experts and of being directly convertible into production rules" [44].

For this reason, we selected a Decision-Tree classifier to evaluate the performance in terms of classification accuracy of all the 63 possible combinations of the six features sets. We limited the maximum depth of the tree to five levels to prevent overfitting and to generate short human-readable rules. Table 3 shows the mean classification accuracy achieved with the combination of the different feature sets.

Considering the mean values, we observed that the greater the number of feature sets involved, the better the classification accuracy of the models. In isolation, the digital level achieved the highest accuracy of 56.9%. The combination of the digital and the lexical levels increased the classification accuracy by 2%. The maximum accuracy of 59.2% was yielded by adding the syntactic feature set to the previous combination. From Table 3, it can be observed that the incorporation of the morphological, pragmatics-discursive and orthographic features did not lead to an increase in accuracy.

**Table 3.** Features and classification accuracy.

| Features | CA |
|---|---|
| D | 56.9 |
| D + L | 58.9 |
| D + L + S | 59.2 |
| D + L + S + M | 59.0 |
| D + L + S + M + P | 59.0 |
| D + L + S + M + P + O | 58.7 |

In Figure 1, we partially reproduce the tree belonging to the DLS model. As it can be observed, digital features, such as the GIF/words ratio and lexical features, such as the frequency of words with appreciative suffixes occupy the first levels of the tree.

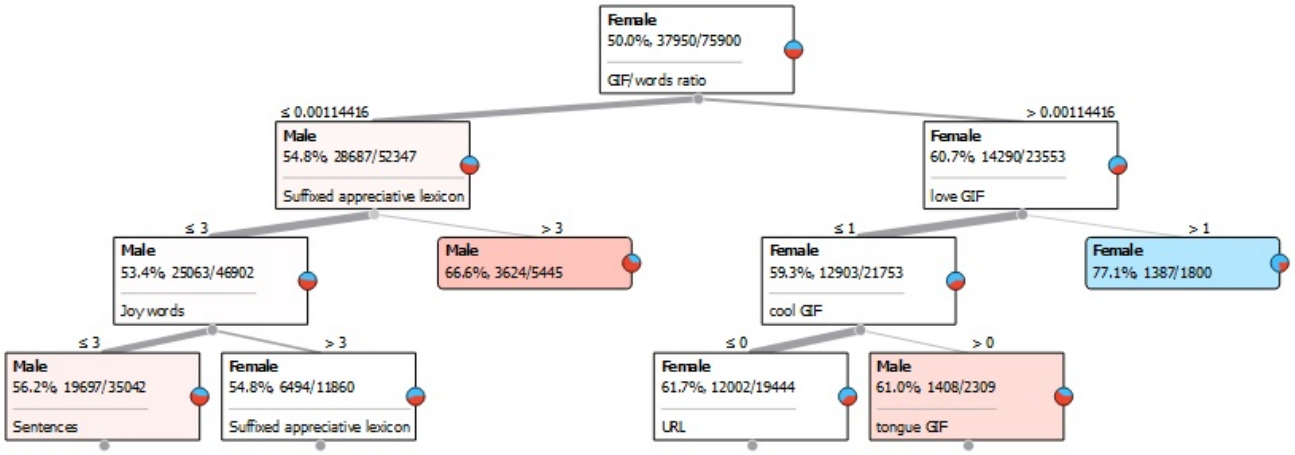


**Figure 1.** The first four levels of the DLS tree.

## 4. Results

To better understand the significance of the features and to be able to trace a sociolinguistic explanation, we ranked the features according to the values provided by the attribute feature_importances_ of the *scikit-learn* (https://scikit-learn.org/stable/, accessed on 17 December 2021) library, which computed the feature importance as the mean and standard deviation of accumulation of the impurity decrease within each tree.

### *4.1. Orthography*

The tree-based classifier discarded 16 out of 29 orthographic features. As shown in Table 4, female-authored texts contained significantly more ellipsis points ($M$ = 3.764), repetition of exclamation marks ($M$ = 0.138), upper-case characters ($M$ = 56.448), and repetition of vowels ($M$ = 0.066), whereas male-authored texts included significantly more numeric characters ($M$ = 6.418), lower-case characters ($M$ = 1139.409), and punctuation marks, such as dashes ($M$ = 4.660), commas ($M$ = 15.430), double quotation marks ($M$ = 1.336), parentheses ($M$ = 1.828), and full-stops ($M$ = 7.771).

Some of these sociolinguistics patterns are consistent with previous work: refs. [15,27] detected that females included more ellipsis points in their messages; ref. [27] concluded that females were more than twice as likely to repeat exclamation marks than male users; refs. [15,45] found that character flooding was among the most informative features in the female category, and [46] indicated that females wrote more upper-case characters as expressiveness markers. In contrast, numeric characters have been previously correlated with male-authored texts in [47,48].

Although frequencies of punctuation marks have been considered as features on AP models [49], as far as we know, gender studies tended to focus mainly on question and exclamation marks [50] or on non-standard orthography [51], and thus further research needs to be conducted on other punctuation marks, such as dashes, commas, double quotation marks, parentheses, and full-stops, which were correlated with male-authored messages according to our results. However, refs. [52,53] found that females used, on average, more punctuation marks.

**Table 4.** Orthographic feature importance and mean values.

| Feature | Importance | Female | Male |
|---|---|---|---|
| Ellipsis points | 0.382 | 3.764 | 2.841 |
| Numeric characters | 0.154 | 4.285 | 6.418 |
| Repetition of exclamation marks | 0.145 | 0.318 | 0.195 |
| Dashes | 0.102 | 3.760 | 4.660 |
| Commas | 0.079 | 14.115 | 15.430 |
| Consonants | 0.050 | 584.344 | 645.232 |
| Double quotation marks | 0.023 | 1.073 | 1.336 |
| Upper case | 0.018 | 56.448 | 52.007 |
| Duplication of exclamation marks | 0.014 | 0.138 | 0.080 |
| Parentheses | 0.012 | 1.264 | 1.828 |
| Lower case | 0.011 | 1026.719 | 1139.409 |
| Repetition of vowels | 0.007 | 0.066 | 0.036 |
| Full stops | 0.005 | 6.884 | 7.771 |

### 4.2. Morphology

Regarding morphology, 19 out of 30 morphological features were discarded by the classifier. As shown in Table 5, female-authored posts included more personal pronouns ($M$ = 7.789), personal ($M$ = 22.458) and non-personal ($M$ = 9.434) verbs, possessive determiners ($M$ = 5.360), coordinating conjunctions ($M$ = 9.041), and demonstrative pronouns ($M$ = 0.492). Male-authored messages contained more prepositions ($M$ = 20.996), numeral determiners ($M$ = 2.689), demonstrative determiners ($M$ = 1.573), definite articles ($M$ = 10.860), and nouns ($M$ = 30.151).

These morphological patterns have already been detected in previous work. Several works have reported the use of personal pronouns and verbs by females [15,26,30,54,55]. Argamon [56] also found that females used more conjunctions than males. On the other hand, male-authored texts contained, in general, more determiners, prepositions and nouns, as shown in [26,30,48,57].

**Table 5.** Morphological feature importance and mean values.

| Feature | Importance | Female | Male |
|---|---|---|---|
| Personal pronouns | 0.305 | 7.789 | 6.215 |
| Prepositions | 0.241 | 20.418 | 20.996 |
| Numeral determiners | 0.174 | 1.931 | 2.689 |
| Demonstrative determiners | 0.072 | 1.528 | 1.573 |
| Non-personal verbs | 0.057 | 9.434 | 8.304 |
| Personal verbs | 0.055 | 22.458 | 20.037 |
| Possessive determiners | 0.033 | 5.360 | 4.535 |
| Definite articles | 0.024 | 10.318 | 10.860 |
| Coordinating conjunctions | 0.024 | 9.041 | 8.597 |
| Demonstrative pronouns | 0.008 | 0.492 | 0.488 |
| Nouns | 0.007 | 29.551 | 30.151 |

### 4.3. Lexicon

Regarding the lexical level, the classifier removed 21 out of 33 lexical features. It can be seen in Table 6 that, from a lexical perspective, females included more emotive terms ($M$ = 10.918) and, specifically, joy- ($M$ = 5.585) and sadness-related ($M$ = 2.680) words. They also employed more mitigating lexical elements ($M$ = 5.385) in order to attenuate their statements. Male-authored messages contained more approximators or numerical hedges, derived words with appreciative affixes ($M$ = 5.138) and, specifically, suffixed words ($M$ = 1.247).

Finally, males exhibited a higher letters/words ratio ($M$ = 4.367), in fact, they wrote more words over six characters ($M$ = 50.317), and they presented a higher lexical diversity ($M$ = 0.758).

Tannen [58] concluded that females tended to have a more supportive orientation. For this reason, they used more attenuated assertions and mitigating lexical elements [59]. This finding is supported by our computational analysis. However, our results differ from well-established conclusions regarding the use of diminutives and suffixed words, since, according to [59–62], females included more diminutives in their texts. Finally, our results also differ from [26], who found that vocabulary richness was associated with female-authored texts.

**Table 6.** Lexical feature importance and mean values.

| Feature | Importance | Female | Male |
|---|---|---|---|
| Joy words | 0.449 | 5.585 | 4.923 |
| Suffixed appreciative lexicon | 0.217 | 0.913 | 1.247 |
| Ratio letters/words | 0.172 | 4.317 | 4.367 |
| Derived appreciative lexicon | 0.040 | 4.215 | 5.138 |
| Approximators | 0.026 | 0.062 | 0.092 |
| Words over six characters | 0.024 | 43.957 | 50.317 |
| Sadness words | 0.019 | 2.680 | 2.566 |
| Lexical diversity | 0.007 | 0.742 | 0.758 |
| TTR Lemma | 0.005 | 0.151 | 0.156 |
| Emotive lexicon | 0.004 | 10.918 | 10.194 |
| Mitigating lexicon | 0.004 | 5.385 | 4.994 |

### 4.4. Syntax

Regarding the syntactic level, the classifier discarded 20 out of 29 syntactic features. According to the results shown in Table 7, females wrote more sentences ($M$ = 3.854), and their messages contained more adverbial modifiers ($M$ = 15.101), open clausal complements ($M$ = 5.361), clausal complements ($M$ = 4.310), and subordinate structures ($M$ = 29.427). On the other hand, males wrote longer sentences ($M$ = 74.049), and their texts included more nominal ($M$ = 12.441) and adjectival ($M$ = 14.169) modifiers as well as flat multiword expressions ($M$ = 8.442).

Thomson [16] also found that males wrote slightly longer sentences than females. Regarding the syntactic dependencies, it should be noted that automatic gender detection studies have focused on sequences of dependencies tags, instead of isolated dependencies. Ref. [63] extracted individual dependency relations; however, they did not provide the distribution of the syntactic features in relation to gender, and therefore we cannot perform a comparative analysis of the results.

**Table 7.** Syntactic feature importance and mean values.

| Feature | Importance | Female | Male |
|---|---|---|---|
| Sentences | 0.354 | 3.854 | 3.408 |
| Nominal modifier | 0.340 | 9.750 | 12.441 |
| Adverbial modifier | 0.170 | 15.101 | 14.673 |
| Sentences length | 0.041 | 71.554 | 74.049 |
| Flat multiword expression | 0.032 | 6.507 | 8.442 |
| Open clausal complement | 0.021 | 5.361 | 5.036 |
| Clausal complement | 0.018 | 4.310 | 4.208 |
| Adjectival modifier | 0.016 | 11.800 | 14.169 |
| Subordination | 0.008 | 29.427 | 29.321 |

### 4.5. Digital Features

At the digital level, the classifier discarded 6 out of 16 digital features. As shown in Table 8, messages posted by females exhibited a higher GIF/words ratio ($M$ = 0.023). In fact, female-authored texts contained more emoticons on average ($M$ = 1.301) and, specifically, more love-related emoticons ($M$ = 0.213). In addition, they also shared more images ($M$ = 0.097) in their posts. In contrast, male-authored texts included more URLs ($M$ = 0.253) and more *cool* emoticons ($M$ = 0.063).

Our findings are also consistent with previous sociolinguistic work. [15,27,64] found that females tended to reinforce their messages with non-verbal communication elements, such as emoticons in order to express emotion. Moreover, females' preference for love-related emoticons has already been indicated by [65]. Regarding the embedded images, ref. [66] found that females shared more photos than males. On the other hand, refs. [22,28] also detected that male-authored messages included more URLs. This pattern has been frequently related to the preference of male users for the informational dimension of communication. However, a qualitative analysis is necessary in order to provide empirical evidence for this conclusion.

**Table 8.** Digital feature importance and mean values.

| Feature | Importance | Female | Male |
|---|---|---|---|
| GIF/words ratio | 0.530 | 0.023 | 0.015 |
| love GIF | 0.213 | 0.213 | 0.065 |
| cool GIF | 0.128 | 0.034 | 0.063 |
| JPG | 0.031 | 0.097 | 0.065 |
| w00t GIF | 0.029 | 0.045 | 0.051 |
| hug GIF | 0.024 | 0.037 | 0.013 |
| URL | 0.021 | 0.185 | 0.253 |
| tongue GIF | 0.015 | 0.117 | 0.071 |
| GIF | 0.004 | 1.301 | 0.795 |
| inlove GIF | 0.003 | 0.043 | 0.017 |

### 4.6. Pragmatic and Discourse

Finally, regarding the pragmatic-discursive level, the classifier did not consider 52 out of 62 pragmatics-discursive features. As shown in Table 9, female-authored texts contained more exclamative sentences ($M$ = 1.201), personal ($M$ = 13.273) and temporal ($M$ = 2.181), deictics, expressions of gratitude ($M$ = 0.174), and lexical presuppositions ($M$ = 1.958). In addition, their messages contained more tabulations ($M$ = 3.823) and line breaks ($M$ = 16.480) Male-authored texts contained more existential presuppositions with proper names and nouns phrases with defined interpretation, spatial deictics ($M$ = 3.390).

Our results are consistent with previous work, such as those reported in [15,55] who found that females wrote more exclamative sentences and [21] who detected that females used more contextual or deictic words than males. Moreover, traditionally, politeness has been related to the female gender [67]. As far as we know, previous gender detection models did not include pragmatics presuppositions as features. Therefore, further research is needed to draw strong sociolinguistic conclusions in this regard.

**Table 9.** Pragmatics-discursive feature importance and mean values.

| Feature | Importance | Female | Male |
|---|---|---|---|
| Exclamative sentences | 0.266 | 1.201 | 0.850 |
| Existential presuppositions proper names | 0.259 | 4.438 | 6.227 |
| Personal deixis | 0.161 | 13.273 | 10.841 |
| Spatial deixis | 0.123 | 3.225 | 3.390 |
| Ex. pr. det. phrases with defined interpretation | 0.058 | 10.428 | 12.064 |
| Temporal deixis | 0.028 | 2.181 | 1.812 |
| Tabulations | 0.027 | 3.823 | 3.152 |
| Line breaks | 0.024 | 18.515 | 16.480 |
| Gratitude expressions | 0.022 | 0.174 | 0.142 |
| Existential presuppositions | 0.018 | 17.642 | 20.967 |
| Lexical presuppositions | 0.014 | 1.958 | 1.744 |

## 5. Conclusions

Is it possible to identify the author's gender from a given text? This is the question we investigated in this paper. The rise of social media, the fact that text is the most prevalent media type used in our digital activity, and people's tendency to hide their identity on social media platforms have shown the potential of authorship profiling. In this paper, we focused on gender identification, a subtask of the authorship profiling problem that aims at determining the gender of the author of a given text, and this has revealed an interesting research area that could benefit forensics, marketing analysis, advertising, sociolinguistics, etc.

Classifying the gender of a person based on short messages is a difficult task since we have to deal with short length and multi-content text. The main goal of our research is to determine sociolinguistic patterns related to gender that could improve automatic gender detection in author profiling tasks. In fact, to claim that gender identification is possible means to assume that men and women generally use different classes of language and that we can identify linguistic features that indicate gender. However, identifying a set of features that indicate gender is still an open research problem.

In this paper, Machine-Learning techniques were used to analyze a Spanish social media corpus in order to obtain linguistic patterns to design a computational model for gender detection. A tree-based computational model made up of 198 features was proposed. Unlike previous work, we handled a reduced set of features that covered a wide linguistic spectrum (orthographic, morphological, lexical, syntactic, pragmatic-discursive, and digital).

A decision tree classifier to evaluate the performance of feature combinations was implemented. Experiments on our corpus indicated an accuracy up to 59.2% in identifying gender. Our experiments also indicated that digital, lexical, and syntactic features were significant gender discriminators. Despite that our model reached less accuracy than other models, we obtained competitive results handling less than 1% of the number of features used in more accurate models.

In modeling our problem, we made a decision regarding the trade-off between model accuracy and model interpretation. In our work, the interpretability of the model was prioritized over the classification accuracy. Although opaque methods in gender detection generally obtain higher accuracy, we chose to pay a penalty in terms of predictive performance when selecting an interpretable model that allows for human/linguistic understanding.

To find linguistic patterns correlated to gender is a common interest in gender identification tasks (within the area of author profiling) and in the field of sociolinguistics. Therefore, the close collaboration between researchers in these two areas can produce better results that benefit both research fields. This is why we claim that the interdisciplinarity is a must when dealing with gender, language, and computation.

As future research avenues, we will replicate this computational analysis on other PAN-AP datasets in order to validate some sociolinguistic patterns. In addition, we will conduct a qualitative analysis to fully understand some of the quantitative results presented in this study.

**Author Contributions:** Conceptualization, D.M.S., A.M., and M.D.J.L.; methodology, D.M.S., A.M., and M.D.J.L., formal analysis, D.M.S.; investigation, D.M.S.; writing–original draft preparation, D.M.S.; writing–review and editing, D.M.S., A.M., and M.D.J.L.; supervision, A.M. and M.D.J.L. All authors have read and agreed to the published version of the manuscript

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The part of the data that supports the findings of this study is available on request from the corresponding authors.

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

## References

1. Bruns, A. User-Generated Content. In *The International Encyclopedia of Communication Theory and Philosophy*; Wiley Online Library: Hoboken, NJ, USA, 2016; pp. 1–5.
2. Lazer, D.; Pentland, A.; Adamic, L.; Aral, S.; Barabási, A.; Brewer, D.; Christakis, N.; Contractor, N.; Fowler, J.; Gutmann, M.; et al. Computational Social Science. *Science* **2009**, *323*, 721–723. [CrossRef] [PubMed]
3. Ochoa, X.; Duval, E. Quantitative analysis of user-generated content on the Web. In Proceedings of the WebEvolve2008: Web Science Workshop at WWW2008, Beijing, China, 22 April 2008; pp. 1–8.
4. Argamon, S.; Koppel, M.; Pennabaker, J.W.; Schler, J. Automatically profiling the author of an anonymous text. *Commun. ACM* **2009**, *52*, 119–123. [CrossRef]
5. Biber, J.K.; Doverskipe, D.; Baznik, D.; Cober, A.; Ritter, B.A. Sexual Harassment in Online Communications: Effects of Gender and Discourse Medium. *CyberPsychol. Behav.* **2002**, *5*, 33–42. [CrossRef] [PubMed]
6. Krysowski, E.; Tremewan, J. *Anonimity, Social Norms, and Online Harassment*; Universität Wien: Vienna, Austria, 2015.
7. Bugueño, M.; Mendoza, M. Learning to detect online harassment on Twitter with the transformer. In *Machine Learning and Knowledge Discovery in Databases*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 1–10.
8. Mukhopadhyay, D.; Mishra, K.; Mishra, K.; Tiwari, L. Cyber Bullying Detection Based on Twitter Dataset. In *Machine Learning for Predictive Analysis*; Springer: Singapore, 2020; pp. 87–94. [CrossRef]
9. Nini, A. Developing forensic authorship profiling. *Lang. Law* **2018**, *5*, 38–58.
10. Shen, A. Recommendations as personalized marketing: Insights from customer experiences. *J. Serv. Mark.* **2014**, *28*, 414–427. [CrossRef]
11. Sun, X.; Wiedenbeck, S.; Chintakovid, T.; Zhang, Q. Gender talk: Differences in interaction style in CMC. In Proceedings of the 11th IFIP TC 13 International Conference on Human-Computer Interaction, Rio de Janeiro, Brazil, 10–14 September 2007; pp. 215–218. [CrossRef]
12. Aljohani, T.; Cristea, A.I. Learners Demographics Classification on MOOCs During the COVID-19: Author Profiling via Deep Learning Based on Semantic and Syntactic Representations. *Front. Res. Metrics Anal.* **2021**, *6*, 1–17. [CrossRef]
13. Nguyen, D.; Doğruöz, A.S.; Rosé, C.P.; de Jong, F. Computational Sociolinguistics: A Survey. *Comput. Linguist.* **2016**, *42*, 537–593. [CrossRef]
14. Santosh, K.; Bansal, R.; Shekhar, M.; Varma, V. Author Profiling: Predicting Age and Gender from Blogs—Notebook for PAN at CLEF 2013. In Proceedings of the CLEF 2013 Labs and Workshops, Notebook Papers, CEUR Workshop, Padua, Italy, 22–23 September 2013.
15. Bamman, D.; Eisenstein, J.; Schnoebelen, T. Gender identity and lexical variation in social media. *J. Socioling.* **2014**, *18*, 135–160. [CrossRef]
16. Thomson, R.; Murachver, T. Predicting gender from electronic discourse. *Br. J. Soc. Psychol.* **2001**, *40*, 193–208. [CrossRef]
17. Singh, S. A Pilot Study on Gender Differences in Conversational Speech on Lexical Richness Measures. *Lit. Linguist. Comput.* **2001**, *16*, 251–264. [CrossRef]
18. Corney, M.; De Vel, O.; Anderson, A.; Mohay, G. Gender-preferential text mining of e-mail discourse. In Proceedings of the 18th Annual Computer Security Applications Conference, Washington, DC, USA, 9–13 December 2002; pp. 282–289. [CrossRef]
19. Koppel, M.; Argomon, S.; Shimoni, A.R. Automatically categorizing written texts by author gender. *Lit. Linguist. Comput.* **2002**, *17*, 401–412. [CrossRef]

20. Boulis, C.; Ostendorf, M. A quantitative analysis of lexical differences between genders in telephone conversations. In Proceedings of the 43rd Annual Meetings of the Association for Computational Linguistics, Ann Arbor, MI, USA, 25–30 June 2005; pp. 435–442. [CrossRef]
21. Nowson, J.; Oberlander, J. The identity of bloggers: Openness and gender in personal blogs. In Proceedings of the AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, Stanford, CA, USA, 27–29 March 2006; pp. 163–167.
22. Schler, J.; Koppel, M.; Argamon, S.; Pennebaker, J.W. Effects of age and gender on blogging. In Proceedings of the AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, Stanford, CA, USA, 27–29 March 2006; pp. 199–205.
23. Yan, X.; Yan, L. Gender classification of weblog authors. In Proceedings of the AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, Stanford, CA, USA, 27–29 March 2006; pp. 228–230.
24. Goswami, S.; Sarkar, S.; Rustagi, M. Stylometric analysis of bloggers' age and gender. In Proceedings of the 3rd International AAAI Conference, San Jose, CA, USA, 17–20 May 2009; pp. 214–217.
25. Mukherjee, A.; Liu, B. Improving gender classification of blog authors. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Cambridge, MA, USA, 9–11 October 2010; pp. 207–217. [CrossRef]
26. Otterbacher, J. Inferring gender of movie reviewers: Exploiting writing style, content and metadata. In Proceedings of the 19th ACM International Conference on Information and Knowledge Management, Toronto, ON, Canada, 26–30 October 2010; pp. 369–378. [CrossRef]
27. Rao, D.; Yarowsky, D.; Shreevats, A.; Gupta, M. Classifying latent user attributes in Twitter. In Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents, Toronto, ON, Canada, 30 October 2010; pp. 37–44. [CrossRef]
28. Burger, J.D.; Henderson, J.; Kim, G.; Zarrella, G. Discriminating gender on Twitter. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Edinburgh, UK, 27–31 July 2011; pp. 1301–1309.
29. Fink, C.; Kopecky, K.; Morawski, M. Inferring gender from the content of tweets: A region specific example. In Proceedings of the 6th International AAAI Conference on Web and Social Media, Dublin, Ireland, 4–7 June 2012; Volume 6, pp. 459–462.
30. Ciot, M.; Sonderegger, M.; Ruths, D. Gender inference of Twitter users in non-English contexts. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, WA, USA, 18–21 October 2013; pp. 1136–1145.
31. Alrifai, K.; Rebdawi, G.; Ghneim, N. Arabic Tweeps Gender and Dialect Prediction—Notebook for PAN at CLEF 2017. In Proceedings of the CLEF 2017 Labs and Workshops, Notebook Papers, CEUR Workshop, Dublin, Ireland, 11–14 September 2017.
32. Manna, R.; Pascucci, A.; Monti, J. Gender detection and stylistic differences and similarities between males and females in a dream tales blog. In Proceedings of the 6th Italian Conference on Computational Linguistics (CLiC-it 2019), Bari, Italy, 13–15 November 2019.
33. Park, S.; Woo, J. Gender Classification Using Sentiment Analysis and Deep Learning in a Health Web Forum. *Appl. Sci.* **2019**, *9*, 1249. [CrossRef]
34. Safara, F.; Mohammed, A.S.; Yousif Potrus, M.; Ali, S.; Tho, Q.T.; Souri, A.; Janenia, F.; Hosseinzadeh, M. An Author Gender Detection Method Using Whale Optimization Algorithm and Artificial Neural Network. *IEEE Access* **2020**, *8*, 48428–48437. [CrossRef]
35. Kowsari, K.; Heidarysafa, M.; Odukoya, T.; Potter, P.; Barnes, L.E.; Brown, D.E. Gender detection on social networks using ensemble Deep Learning. In Proceedings of the Future Technologies Conference (FTC), San Francisco, CA, USA, 5–6 November 2020; pp. 346–358. [CrossRef]
36. Sharma, D.J.; Dutta, S.; Bora, D.J. REGA: Real-time emotion, gender, age detection using CNN—A review. In Proceedings of the 2020 International Conference on Research in Management & Technovation (ACSIS, 2020), Nagpur, India, 5–6 December 2020; pp. 115–118. [CrossRef]
37. Sumi, T.A.; Hossain, M.S.; Islam, R.U.; Andersson, K. Human Gender Detection from Facial Images Using Convolution Neural Network. In *Applied Intelligence and Informatics*; Springer International Publishing: Cham, Switzerland, 2021; pp. 188–203.
38. Krishna, D.N.; Amrutha, D.; Sai Sumith, R.; Anudeepa, A.; Prabhu Aashish, G.; Triveni, B.J. Language Independent Gender Identification from Raw Waveform Using Multi-Scale Convolutional Neural Networks. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6559–6563. [CrossRef]
39. Rangel, F.; Rosso, P.; Koppel, M.; Stamatatos, E.; Inches, G. Overview of the Author Profiling Task at PAN 2013. In Proceedings of the CLEF 2013 Labs and Workshops, Notebook Papers, CEUR Workshop, Valencia, Spain, 23–26 September 2013.
40. Neal, T.; Sundararajan, K.; Fatima, A.; Yan, Y.; Xiang, Y.; Woodard, D. Surveying Stylometry Techniques and Applications. *ACM Comput. Surv.* **2017**, *50*, 1–36. [CrossRef]
41. Rangel, I.D.; Sidorov, G.; Guerra, S.S. Creation and evaluation of a dictionary tagged with emotions and weighted for Spanish. *Onomazein* **2014**, *29*, 31–46.
42. Loyola-González, O. Black-Box vs. White-Box: Understanding Their Advantages and Weaknesses From a Practical Point of View. *IEEE Access* **2019**, *7*, 154096–154113. [CrossRef]
43. Song, Y.Y.; Lu, Y. Decision tree methods: Applications for classification and prediction. *Shanghai Arch. Psychiatry* **2015**, *27*, 130–135.
44. Almuallim, H.; Kaneda, S.; Akiba, Y. Development and Applications of Decision Trees. *Expert Syst.* **2002**, *1*, 53–77.

45. Verhoeven, B.; Škrjanec, I.; Pollak, S. Gender profiling for Sloven Twitter communication: The influence of gender marking, content and style. In Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing, Valencia, Spain, 4 April 2017; pp. 119–125. [CrossRef]
46. Parking, R. Gender and Emotional Expressiveness: An Analysis of Prosodic Features in Emotional Expression. *Griffith Work. Pap. Pragmat. Intercult. Commun.* **2012**, *5*, 46–54.
47. Newman, M.L.; Groom, C.J.; Handelman, L.D.; Pennebaker, J.W. Gender differences in language use: An analysis of 14,000 text samples. *Discourse Process.* **2008**, *45*, 211–236. [CrossRef]
48. Hosseini, M.; Tammimy, Z. Recognizing users gender in social media using linguistic features. *Comput. Hum. Behav.* **2016**, *56*, 192–197. [CrossRef]
49. Rangel, F.; Rosso, P. Use of language and author profiling: Identification of gender and age. In Proceedings of the Nautral Language Processing and Cognitive Science, Marseille, France, 15–16 October 2013; pp. 177–186.
50. Waseleski, C. Gender and the Use of Exclamation Points in Computer-Mediated Communication: An Analysis of Exclamations Posted to Two Electronic Discussion Lists. *J. Comput.-Mediat. Commun.* **2006**, *11*, 1012–1024. [CrossRef]
51. Zelenkauskaite, A.; Herring, S.C. Gender encoding of typographical elements in Lithuanian and Croatian IRC. In *Cultural Attitudes Towards Technology and Communication 2006: Proceedings of the Fifth International Conference on Cultural Attitudes towards Technology and Communication, Tartu, Estonia, 28 June–1 July 2006*; Murdoch University Press: Murdoch, Australia, **2006**.
52. Ling, R. The Sociolinguistics of SMS: An Analysis of SMS use by a random sample of Norwegians. In *Mobile Communication and the Recognition of the Social Sphere*; Ling, R., Pederson, P., Eds.; Springer: London, UK, 2005; pp. 335–350.
53. Al Rousan, R.M.; Abd Aziz, N.H.; Christopher, A.A. Gender differences in the typographical features used in the text messaging of young Jordanian undergraduates. In Proceedings of the International Conference on Languages, Literature and Linguistics, Dubai, United Arab Emirates, 28–30 December 2011.
54. Schwartz, H.A.; Eichstaedt, J.C.; Kern, M.L.; Dziurzynski, L.; Ramones, S.M.; Agrawal, M.; Shah, A.; Kosinski, M.; Stillwell, D.; Seligman, M.E.P.; et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE* **2013**, *8*, e73791. [CrossRef] [PubMed]
55. Gianfortoni, P.; Adamson, D.; Rosé, C.P. Modeling of stylistic variation in social media with stretchy patterns. In Proceedings of the 1st Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties, Edinburgh, UK, 31 July 2011; pp. 49–59.
56. Argamon, S.; Koppel, M.; Pennebaker, J.W.; Schler, J. Mining the blogosphere: Age, gender and the varieties of self-expression. *First Monday* **2007**, *12*. [CrossRef]
57. Johannsen, A.; Hovy, D.; Søggard, A. Cross-lingual syntactic variation over age and gender. In Proceedings of the 19th Conference on Computational Language Learning, Beijing, China, 30–31 July 2015; pp. 103–112. [CrossRef]
58. Tannen, D. *You Just Don't Understand: Men and Women in Conversation*; Ballantine: New York, NY, USA, 1990.
59. Lakoff, R. Language and Woman's Place. *Lang. Soc.* **1973**, *2*, 45–80. [CrossRef]
60. García Mouton, P. *Cómo Hablan las Mujeres*; Arco Libros: Madrid, Spain, 1999.
61. García Mouton, P. *Así Hablan las Mujeres. Curiosidades y Tópicos del Uso Femenino del Lenguaje*; La Esfera de los Libros: Madrid, Spain, 2003.
62. Silva-Corvalán, C. *Sociolingüística: Teoría y Análisis*; Editorial Alhambra: Madrid, Spain, 1989.
63. Soler-Company, J.; Wanner, L. On the role of syntactic dependencies and discourse relations for author gender identification. *Pattern Recognit. Lett.* **2018**, *105*, 87–95. [CrossRef]
64. Witmer, D.F.; Katzman, S.L. On-Line Smiles: Does Gender Make a Differnece in the Use of Graphic Accents? *J. Comput.-Mediat. Commun.* **1997**, *2*, JCMC244. [CrossRef]
65. Chen, Z.; Lu, X.; Ai, W.; Li, H.; Mei, Q.; Liu, X. Through a Gender Lens: Learning Usage Patterns of Emojis from Large-Scale Android Users. In Proceedings of the 2018 World Wide Web Conference, Lyon, France, 23–27 April 2018; pp. 763–772. [CrossRef]
66. Mendelson, A.L.; Papacharissi, Z. Look at us: Collective narcissism in college student Facebook photo galleries. In *The Networked Self: Identity, Community and Culture on Social Network Site*; Papacharissi, Z., Ed.; Taylor & Francis: Hoboken, NJ, USA, 2010; pp. 251–273.
67. Holmes, J. *Women, Men and Politeness*; Routledge: London, UK, 1995.