Machine Learning Methods for Automatic Gender Detection

Damián Morales Sánchez

GRLMC, Universitat Rovira i Virgili, Av. Catalunya, 35,
Tarragona, 43002, Spain
damian.morales@estudiants.urv.cat

Antonio Moreno

ITAKA, Universitat Rovira i Virgili, Av. Països Catalans, 26,
Tarragona, 43007, Spain
antonio.moreno@urv.cat

M. Dolores Jiménez López

GRLMC, Universitat Rovira i Virgili, Av. Catalunya, 35,
Tarragona, 43002, Spain
mariadolores.jimenez@urv.cat

Abstract: Automatic gender detection has attracted the attention of many research fields such as forensic linguistics or marketing. Within these areas, gender detection has been approached as a classification problem and, for this reason, supervised Machine Learning algorithms such as Naïve Bayes, Logistic Regression and Support Vector Machines, among others, have been employed. The latter algorithm has exhibited a better performance on gender detection. In recent years, with the development of Deep Learning methods, various neural networks structures such as Convolutional Neural Networks have been designed for gender detection. However, Deep Learning methods have led to a loss in the interpretability of the models. In this article, we review the AI techniques applied on gender detection.

1. Introduction

We live in a digital era and we actively contribute to the construction of this digital reality through our daily interaction with the web. User-generated content has been considered a valuable source of information that has been mined in recent years by areas such as linguistics, marketing, healthcare, and education, among others. Social networks connect millions of users around the world who post updates, share photos, leave comments, and disseminate content posted by others. These digital fingerprints make up huge volumes of complex and multimodal data that have been collected through automatic methods and analyzed with Machine Learning algorithms to discover hidden patterns that can provide useful information to understand social phenomena.

Computational social sciences have applied computational methods and techniques to correlate textual features with social variables. Within this research

agenda, *author profiling* algorithms analyze users' messages to find out personal and social traits such as personality, age, or gender. Beyond author profiling, various research fields have taken interest in *automatic gender detection*, mainly forensic linguistics, marketing, and sociolinguistics.

The exponential growth of the web and social networks such as Instagram, Twitter, Facebook or TikTok, has also led to an increase in violence and harassment in cyberspace. Sexual predators often create a false digital identity to impersonate a teenager and to contact with their victims: "a common characteristic of these digital communities is that it is easy to provide a false name, age, gender and location in order to hide one's true identity, providing criminals such as pedophiles with new possibilities to groom their victims" [1]. Given the relative ease with which personal data can be falsified, the linguistic productions made by a suspected user on the web become a source of information about its true identity. Forensic linguistics traces the profile of a suspected author from the linguistic style and other textual variables of the messages, so that they can constitute solid evidence for legal investigation. The automatic detection of gender is essential in these cases, since gender, along with other sociodemographic information such as age or native language, allows forensic linguistics to reconstruct the real identity of the person who deliberately hides behind a fake profile.

On the other hand, many commercial companies have increased their interest in developing personalization strategies through the automated analysis of the activity and contributions of the users on the Web. With the consolidation of e-commerce, advertising companies have invested in predicting certain personal data of their customers in order to better define their market niche and to be able to personalize more effectively their advertising messages. Various proposals have been developed for predicting customers' gender from their texts and from non-textual data such as catalog viewing data or website traffic [2]. Closely related to marketing, opinion mining companies monitor users' attitudes and feelings regarding products, services or ideas from review websites or forums in order to make decisions of a different nature.

Finally, automatic gender detection can contribute to linguistic disciplines such as sociolinguistics. Sociolinguists are concerned with finding linguistic patterns correlated with social variables, including gender. Conclusions about gender were frequently drawn from face-to-face informal interactions and occasionally from very large samples. The introduction of computational techniques, especially intelligent data analysis and Machine Learning algorithms, allows managing larger samples of speakers to review the linguistic patterns defined by sociolinguistics in relation to gender. These patterns may include linguistic information of different kinds, including lexical, orthographic, syntactic and semantic aspects. The following section presents the evolution of the use of AI techniques in gender detection.

2. Machine Learning Methods for Automatic Gender Detection

Automatic gender detection is a key task for computational sociolinguistics. This research field, based on the interdisciplinarity between sociolinguistics and computer science, aims to extract linguistic patterns correlated with social variables such as gender or age from the implementation of Artificial Intelligence methods such as Machine Learning algorithms.
Within this field, gender detection has been treated as a binary classification task and researchers have focused on supervised Machine Learning classification algorithms. Four different stages may be identified in the evolution of this area.

2.1. *First stage (2001-2005)*

Initial studies on automatic gender detection worked with formal texts and reduced samples. In this first stage, discriminant analysis was used to automatically predict gender. In 2001, Thomson and Murachver examined gender-preferential language styles in informal electronic discourse [3] and Singh conducted a study on gender differences in recorded conversations [4]. In both studies, the reported accuracy was around 90%. However, they handled very small samples: 35 and 30 participants, respectively.

A year later, Corney *et al*. used a Support Vector Machine (SVM) to analyze an e-mail dataset made up of 8,820 e-mail documents from 342 authors [5]. They reported 70.2% F1-score. Previously, SVMs had already been employed in other tasks such as spam e-mail categorization or authorship attribution. Koppel, Argamon and Shimoni designed a learning method based on the Exponential Gradient algorithm to examine a subset of documents extracted from the British National Corpus [6]. With this method, they reached 80% of accuracy.

In 2005, Boulis and Ostendorf experimented with various Machine Learning algorithms, namely Rocchio, Naïve Bayes, Maximum Entropy and SVM [7]. This latter algorithm exhibited better performance. They reported 92.5% of accuracy regarding gender detection on telephone conversations with SVM.

2.2. *Second stage (2006-2010)*

In a second stage, automatic gender detection focused on analyzing informal texts belonging to the blogosphere. In addition, studies tried to predict other sociodemographic variables. In this way, researchers were not only interested in detecting the gender of the authors, but also their age or origin. In 2006, Nowson and Oberlander also applied a SVM on their analysis conducted on blogs [8]. Specifically, their corpus consisted of 71 weblogs made up of 410,000 words. They reported 91.5% of accuracy on gender detection. Schler *et al*. examined 71,000 blogs using the Multi-Class Real Winnow (MCRW) algorithm [9]. They

implemented MCRW to classify blogs according to author gender and age. With this learning method, they achieved 80.1% of accuracy. Yan and Yan also experimented with weblogs [10]. In their study, they analyzed 75,000 blogs with Naïve Bayes, reporting 68% F1-score.

Similar to Schler *et al*., Goswami, Sarkar and Rustagi performed gender and age detection on 20,000 blog entries [11]. They implemented a Naïve Bayes classifier and reached an accuracy of 89.3% on gender detection.

In 2010, Mukherjee and Liu analyzed 3,100 blogs [12]. Although they tested various algorithms, SVM algorithm surpassed the others. Specifically, they reported 88.56% of accuracy using a SVM regression algorithm. Otterbacher applied a Logistic Regression classifier on a movie review corpus, made up of 31,300 reviews, sourced from the Internet Movie Database (IMDb) platform [13]. He reached a classification accuracy of 73.3%.

2.3. *Third stage (2011-2013)*

In a third stage, automatic gender detection studies focused on microblogging and, more specifically, on Twitter. Rao, Yarowsky, Shreevats and Gupta examined gender, age, regional origin and political orientation on Twitter [14]. Their corpus consisted of 1,000 Twitter users. They applied a SVM and achieved 72.33% of classification accuracy.

In 2011, Burger, Henderson, Kim and Zarrella collected a large corpus made up of 184,000 Twitter users [15]. They tested Naïve Bayes, Balanced Winnow2 and SVM algorithms. They reached 75.5% of accuracy with a Winnow2 classifier. Peersman, Daelemans and Van Vaerenbergh conducted their study on 1,537,283 Netlog posts [1]. They reached 66.3% of classification accuracy using a SVM-based classifier.

In 2012, Fink, Kipecky and Morawski collected 78,853 Twitter users [16]. With the implementation of a SVM, they reached 80.6% of classification accuracy.

In 2013, Ciot, Sonderegger and Ruths extracted Twitter texts written in four different languages: French, Indonesian, Turkish, and Japanese [17]. They achieved an accuracy that oscillated between 63% on Japanese to 87% on Turkish using a SVM classifier.

2.4. *Fourth stage (since 2014)*

In the last years, automatic gender detection has incorporated the latest developments in Machine Learning. Recent studies are based on Deep Learning structures. In 2019, Manna, Pascucci and Monti designed a Feed-Forward Neural Network to analyze 56 blogs from the SogniLucidi platform [18]. They reported 77.6% of classification accuracy.

In 2020, Safara *et al.* employed an Artificial Neural Network on the Enron corpus [19]. In addition, they applied the Whale Optimization Algorithm to find

the optimal weights and improve the accuracy of the neural network structure. With this method, they achieved 98% of classification accuracy, outperforming previous work on the Enron corpus. Kowsari *et al*. implemented Deep Neural Networks and Convolutional Neural Networks on the PAN 2017 dataset collected from Twitter [20]. They reported 86.33% of classification accuracy.

Beyond gender detection from textual data, Deep Learning methods have been employed in other kind of data such as visual data [21]. Sharma, Dutta and Bora used Convolutional Neural Networks to detect gender from face images on webcam [22], and Krishna *et al*. used Multi-Scale Convolutional Neural Networks on 175 hours of audio data [23].

2.5. *Machine Learning methods on Author Profiling PAN tasks*

In this later stage, and within the frame of the *Conference and Labs of the Evaluation Forum* (CLEF), the *Uncovering Plagiarism, Authorship and Social Software Misuse* (PAN) association holds an annual congress dedicated to author profiling, among other tasks, in which research teams can join in and participate under the same computational conditions. From 2013 to date, seven editions have addressed gender detection, constituting a benchmark in this task. Although the datasets were mostly collected from Twitter, they also experimented with blogposts, hotel reviews, essays, and even multimodal datasets made up of text and images. The following table shows the algorithms that exhibited a greater performance on the automatic detection of gender in Spanish. As it can be observed, SVM predominated in most of the editions.

Table 1. Best Machine Learning algorithms on Spanish gender detection PAN 2013-19 tasks.

Edition	Algorithm	Accuracy (%)
2013	Decision Tree	64.73
2014 (a)	Logistic Regression	68.37
2014 (b)	Logistic Regression	65.56
2014 (c)	Support Vector Machine	58.93
2015	Support Vector Machine	96.59
2016	Support Vector Machine	73.21
2017	Support Vector Machine	83.21
2018	Support Vector Machine	82
2019	Support Vector Machine	81.72

In 2013, the winner extracted style-based features such as punctuation marks frequencies, word unigrams, and Latent Dirichlet Analysis-based topics. In 2014, the PAN organization committee provided participants with three subcorpus made up of: a) blogs, b) tweets, and c) social media posts. The highest accuracy

was reached on the blogs dataset with a bag-of-words model. In 2015, Álvarez Carmona *et al*. combined second order attributes with Latent Semantic Analysis to achieve the maximum classification accuracy ever reached in any PAN edition [24]. Finally, from 2016 to 2019, the winners focused on character and word *n*-gram models. Specifically, in 2016, Gencheva *et al*. combined character trigrams and word unigrams and bigrams with style-based, sentiment and topic features [25], whereas Basile *et al*. [26] and Daneshvar and Inkpen [27] in 2017 and 2018, respectively, extracted character *n*-grams in the range (3, 5) and word unigrams and bigrams. Similarly, in 2019, the best model was based on character and word unigrams, bigrams and trigrams.

3. Discussion

Automatic gender detection has experimented a significant change since its inception. Deep Learning methods have burst into gender detection studies. However, those fields interested in gender detection are also interested in the modelling of gender. For this reason, the implementation of Machine Learning algorithms such as Decision Tree or Naïve Bayes allows researchers to interpret the models unlike the so-called "black-box" algorithms: "Model interpretability of Deep Learning […] has always been a limiting factor for use cases requiring explanations of the features involved in modelling"[28]

Moreover, the difference in accuracy between Deep Learning methods and other traditional Machine Learning algorithms is not always remarkable. For example, the Deep Learning model designed by Kowsari *et al*. [20] presented only 3.12% more accuracy than the SVM model of Basile *et al*. [26] on the same dataset. However, one may wonder whether a classification model based on character n-grams is meaningful from the sociolinguistic point of view.

In conclusion, the implementation of Machine Learning methods should be governed by the premise of interpretability in order to provide social sciences with a solid empirical basis with which to refine and design sociological models and explanations.

References

1. C. Peersman, W. Daelemans and L. Van Vaerenbergh, Predicting age and gender in online social networks, in *Proceedings of the 3rd International Workshop on Search and Mining User-Generated Contents* (ACM, 2011), pp. 37-44, doi:10.1145/2065023.2065035
2. D. Duong, H. Tan and S. Pham, Customer gender prediction based on e-commerce data, in *2016 8th International Conference on Knowledge and Systems Engineering* (KSE, 2016), pp. 91-95, doi: 10.1109/KSE.2016.7758035
3. R. Thomson and T. Murachver, Predicting gender from electronic discourse, *British Journal of Social Psychology* **40**(1) (2001) 193-208, doi:10.1348/014466601164812

4. S. Singh, A pilot study on gender differences in conversational speech on lexical richness measures, *Literary and Linguistic Computing* **16**(3) (2001) 251-264, doi:10.1093/llc/16.3.251
5. M. Corney, O. De Vel, A. Anderson and G. Mohay, Gender-preferential text mining of e-mail discourse, in *18th Annual Computer Security Applications Conference*, (IEEE, 2002), pp. 282-289, doi:10.1109/CSAC.2002.1176299
6. M. Koppel, S. Argamon and A. R. Shimoni, Automatically categorizing written texts by author gender, *Literary and Linguistic Computing* **17**(4) (2002) 401-412, doi:10.1093/llc/17.4.401
7. C. Boulis and M. Ostendorf, A quantitative analysis of lexical differences between genders in telephone conversations, in *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics* (ACL, 2005), pp. 435-442, doi:10.3115/1219840.1219894
8. S. Nowson and J. Oberlander, The identity of bloggers: Openness and gender in personal weblogs, in *Proceedings of the AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs* (AAAI, 2006)
9. J. Schler, M. Koppel, S. E. Argamon and J. W. Pennebaker, Effects of age and gender on blogging, in *Proceedings of the AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs* (AAAI, 2006)
10. X. Yan and L. Yan, Gender classification of weblog authors, in *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs* (AAAI, 2006)
11. S. Goswami, S. Sarkar and M. Rustagi, Stylometric analysis of bloggers' age and gender, in *Proceedings of the 3rd International ICWSM Conference* (ICWSM, 2009), pp. 214-217.
12. A. Mukherjee and B. Liu, Improving gender classification of blog authors, in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (ACL, 2010), pp. 207-217, doi:10.5555/1870658.1870679
13. J. Otterbacher, Inferring gender of movie reviewers: Exploiting writing style, content and metadata, in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management* (ACM, 2010), pp. 369-378, doi:10.1145/1871437.1871487
14. D. Rao, D. Yarowsky, A. Shreevats and M. Gupta, Classifying latent user attributes in Twitter, in *Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents* (ACM, 2010), pp. 37-44, doi:10.1145/1871985.1871993
15. J. D. Burger, J. Henderson, G. Kim and G. Zarrella, Discriminating gender on Twitter, in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (ACL, 2011), pp. 1301-1309.
16. C. Fink, J. Kopecky and M. Morawski, Inferring gender from the content of tweets: A region specific example, in *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media* (AAAI, 2012), pp. 459-462.
17. M. Ciot, M. Sonderegger and D. Ruths, Gender inference of Twitter users in non-English contexts, in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (ACL, 2013), pp. 1136-1145.
18. A. Pascucci, V. Masucci and J. Monti, Computational stylometry and Machine Learning for gender and age detection in cyberbullying texts, in *Proceedings of the 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos* (ACIIW, 2019), pp. 1-6, doi:10.1109/ACIIW.2019.8925101

19. F. Safara, A. S. Mohammed, M. Yousif Potrus, S. Ali, Q. T. Tho, A. Souri, F. Janenia and M. Hosseinzadeh, An author gender detection method using whale optimization algorithm and artificial neural network, *IEEE Access* **8**(1) (2020) 48428-48437, doi:10.1109/ACCESS.2020.2973509
20. K. Kowsari, M. Heidarysafa, T. Odukoya, P. Potter, L. E. Barnes and D. E. Brown, Gender detection on social networks using ensemble Deep Learning, *ArXiv* abs/2004.06518 (2020)
21. M. Hussain, I. Ullah, H. A. Aboalsamh, G. Muhammad, G. Bebis and A. M. Mirza, Gender recognition from face images with dyadic wavelet transform and local binary pattern, *International Journal on Artificial Intelligence Tools* **22**(06) (2013), doi: 10.1142/S021821301360018X
22. D. Jyoti Sharma, S. Dutta and D. Jyoti Bora, 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), pp. 115-118, doi: 10.15439/2020KM18
23. D. N. Kirshna, D. Amrutha, S. S. Reddy, A. Acharya, P. A. Grapati and B. J. Triveni, Language independent gender identification from raw waveform using multi-scale convolutional neural networks, in *ICASSP 2020 – 2020 IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP, 2020), pp. 6559-6563, doi:10.1109/ICASSP40776.2020.9054738
24. M. A. Álvarez Carmona, A. P. López Monroy, M. Montes y Gómez, L. Villaseñor Pineda and H. J. Escalante, INAOE's participation at PAN'15: Author profiling task, in *CLEF 2015 Labs and Workshops, Notebook Papers* (CLEF, 2015)
25. P. Gencheva, M. Boyanov, E. Deneva, P. Nakov, Y. Kiprov, I. Koychev and G. Georgiev, PANcakes team: A composite system of genre-agnostic features for author profiling, in *CLEF 2016 Working Notes. CEUR Workshop Proceedings* (CLEF, 2016)
26. S. Daneshvar and D. Inkpen, Gender identification in Twitter using n-grams and LSA, in *CLEF 2018 Working Notes. CEUR Workshop Proceedings* (CLEF, 2018)
27. A. Basile, G. Dwyer, M. Medvedeva, J. Rawee, H. Haagsma and M. Nissim, N-GrAM: new Groningen author -profiling model, in *CLEF 2017 Working Notes. CEUR Workshop Proceedings* (CLEF, 2017)
28. K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes and D. Brown, Text classification algorithms: A survey, *Information* **10**(4) (2019), doi: 10.3390/info10040150