Artificial Intelligence Research and Development
A. Cortés et al. (Eds.)

doi:10.3233/FAIA220315

Drake or Hen? Machine Learning for Gender Identification on Twitter

Arnault GOMBERT [a,1], and Jesus CERQUIDES [b]
[a] *Citibeats, Barcelona, Spain*
[b] *Artificial Intelligence Research Institute (IIIA), CSIC, Cerdanyola 08193, Spain*

Abstract. Social media offers an invaluable wealth of data to understand what is taking place in our society. However, the use of social media data to understand phenomena occurring in populations is difficult because the data we obtain is not representative and the tools which we use to analyze this data introduce hidden biases on characteristics such as gender or age. For instance, in France in 2021 women represent 51.6% of the population [1] whereas on Twitter they represent only 33.5% of the french users [2]. With such a difference between social networks user demographics and real population, detecting the gender or the age before going into a deeper analysis becomes a priority. In this paper we provide the results of an ongoing work on a comparative study between three different methods to estimate gender. Based on the results of the comparative study, we evaluate future work avenues.

Keywords. machine learning; deep learning; bias; NLP; gender; Twitter

1. Introduction

Social networks provide a rich amount of real time data that social scientists analyze to find actionable insights. For instance [3] investigates Chilean citizens perception of transport, [4] looks at how social media influenced the 2016 presidential campaign in the United States or [5] focuses on emotions and narratives with respect to gender. The inherent bias carried by social networks illustrates the need to identify gender for social scientists when analyzing the networks before drawing any assumption or conclusion.

Twitter has an easy access to their data using their own API[2]. Besides, the community has focused mainly a lot on the Twitter data. For instance [3,5,6,7,8,9,10,11] based their research only on tweets and did not address other social networks. As argued by Steinert-Threlkeld [12]: "Twitter presents an ideal combination of size, international reach, and data accessibility that make it the preferred platform in academic studies". In this work we tackle the problem of gender identification on Twitter.

There are two main research lines dealing with gender identification on Twitter. The first one deals with the capability to identify the gender of an author based fundamentally on the texts generated by the user. This research line is very well summarized by

[1] Corresponding Author: Arnault Gombert, Citibeats, Barcelona, Spain. E-mail:agombert@citibeats.com
[2] https://developer.twitter.com/en/docs/twitter-api

Ikae and Savoy in [13] and has as corner stone the author profiling task at PAM-CLEF competitions [14,15,16,17,18,19,20].

The second one does also use the metadata related to the tweet and to its author that can be obtained via the Twitter API. These tweet and author metadata are used as inputs of the gender identification model. Quite some work has been devoted to identify gender just based on the name since the seminal work of Michael [21]. Among them, we highlight Demographer [22,23] and [24]. Other works [3] tackle the problem with classical machine learning techniques converting *names*, *description* and/or *username* in a *bag-of-words* representation through word or character *n-grams*. We also see a development of gender identification through deep-learning methods, for instance [8,9] added a computer vision component to process the user profile picture to improve results. And [9] processed *names*, *description* and *username* through a sequential model. More recently we saw models with pre-trained BERT-architecture [25] like in [6] to determine the user's gender. And [10] focused on the difference between using classical machine learning and deep learning methods.

One of the main difficulties is to build one model dealing with several languages at the same time, a lot of works [3,4,5,6,10,22,7] focused on one or two languages only. Wang et al. [9] are the first ones to tackle the problem in a multilingual paradigm, processing 32 different languages. Nevertheless, the possibility to evaluate the models and create a benchmark is complicated due to the lack of labeled data in those several languages.

A more fine grained understanding of the typology of Twitter users is provided when we are able to discern between human accounts from accounts coming from organizations as in [26]. This line of work is continued in [9] and we do also focus our work in finding out if users are real humans or institutional account such as a company or official pages.

In this work we present an initial set of results of our efforts to set up a production multilingual gender identification system on Twitter. Our contributions are

(i) an efficient and scalable methodology to create multilingual datasets with Twitter users soft labeling, and
(ii) a baseline evaluation of the quality of three models for gender identification based on this data.

Following this line of work, in the near future, we plan to

(i) open a labeled dataset that could be used as test set to benchmark any gender machine learning model,
(ii) open our model via one API like in [11] or in the HuggingFace platform[27], and
(iii) quantify our model bias regarding gender occupations like in [28] and mitigate it if necessary.

2. Methods

2.1. A methodology for building multilingual Twitter user demographics pseudo labeled datasets

We are interested in designing a methodology that can be used to create *reasonably good* datasets of Twitter users annotated as either `institution`, `man`, or `woman`. The

methodology is required to be easily adaptable to any new language. Labeling data is time consuming and as we have access to a vast amount of unlabeled data on Twitter, we focus on soft-labels to fast annotate a large amount of data. We design a procedure that fits all languages.

The main idea of our procedure is to combine different soft-labelers to create more robust datasets. The first one uses self-reporting info as in [3,9,6]. For instance users that have in their description *father and grandpa* or *official account* are soft-labeled respectively as `man` or `institution`. We provide a list of self-reporting words or expressions that can be considered as self-reporting[3]. The second soft-labeler is inspired from [22], we look, for each language, at a list of first names and its number of occurrences for men and women, then we applied it on Twitter `names` to get an estimator of the gender. For instance in the US we have access to the distribution of gender by name[4]. We also used a third soft-labeler based on dependency parsing and *part-of-speech tagging* to detect if the user *description* if the user is expressing it/her/himself as a person, an institution or a group of persons, for instance like *"I love pancakes"* would never refer to an institution whereas a description like *"Our firm helps ..."* would. Then we combine those soft labels to identify better the user demographic with a majority vote across all soft-labelers. In case of ties, we do not include the data into the final set.

2.2. A first soft-labeled dataset

At first, we applied the procedure on five different languages: *Catalan*, *English*, *French*, *Portuguese* and *Spanish*. We sampled Twitter from 2016 to 2020 to get 3 million users in total. Table 1 describes the demographic labels in our datasets after applying our soft-labelers and balancing our dataset in order to have a 50% of the users being institutions, 25% being men and 25% being women. A rough estimate of the equivalent amount of data Wang et al. [9] would have used for training with only 5 languages instead of 32 would be 2.27M users. Thus, our dataset is about a 5% in terms of size of Wang et al. dataset. Table 2 describes the distribution of data across the different languages. Both tables also include the numbers for the split of the total dataset into a 70% for training and a 30% for testing stratified by language.

Dataset	Total	`organization`	`male`	`female`
Total	106810	53405	26703	26702
Train	71562	35781	17891	17890
Test	35248	17624	8812	8812

Table 1. Distribution of collected data by demographic labels.

2.3. Experimental Design

Our model considers as inputs only three of the metadata texts associated with a Twitter user: its *name*, its *description* and its *username*. And it considers a two variables outputs (y_1, y_2) with $y_1 = 1$ if the user is an institution, and 0 otherwise. As for y_2 we have that

[3] https://drive.google.com/drive/u/0/folders/1JP-xOlwzLU8Ue_jRyXGtPRdWrjRWJ3Eu
[4] https://www.ssa.gov/oact/babynames/limits.html

Dataset	Catalan	English	French	Portuguese	Spanish
Total	6560	28606	11173	13890	46581
Train	4390	19220	7411	9434	31107
Test	2170	9386	3762	4456	15474

Table 2. Distribution of collected data across languages

$y_2 = 1$ if the user is a woman, and $y_2 = 0$ if the user is a man and otherwise it does not matter.

We created three different models. The objective is to explore the capacities of a baseline model without machine learning, a classical machine learning approach with a *bag-of-words* representation and a deep learning model in the line of the one presented by Wang et al. [9].

The name may be a strong factor for defining the final gender. But as we want our model to also learns from features present in the description we would like to attenuate the importance given to the name. In order to smooth the name occurrence in the training set we decided to mask the name with a probability inversely proportional to its frequency in the training set. Thus the most frequent name such as Thomas or Juliet will be learnt but the model will also focus more on the description. Furthermore in order to avoid overfitting from the gender masks we used in soft-labels, we decided to mask them with a probability of 80% the gendered marks we have identified with our soft-labelers to avoid overfitting. The loss functions are also adapted as in [29] to consider the second output if and only if the observation is a human.

2.3.1. Baseline model

The baseline model is based on an *a priori* gender marked words such as *mistress* or *waiter* to identify gender. To detect if the user is a human being, we look if the `name` of the user contains a first name from the ones we gathered with the gender distribution in each language for our second soft-labelers. If the `name` does not contain any of the first name then we considered the user as an institution otherwise a human being. Then to differentiate the gender we used the baseline created in [30]. This baseline uses weights from a regression model to detect gender on social networks. We applied this regression on the concatenated `name` and `description`.

2.3.2. Machine learning model

The second model is based on a simple pipeline combining a *term-frequency times inverse document-frequency* (TF-IDF) representation and a logistic regression classifier. The TF-IDF looks at *unigrams* and *bigrams*. It has as parameters a list of predefined stop-words in each language and considers only tokens appearing at least twice in the training set. The logistic regression had a *l1-regularilizer* penalty.

2.3.3. Deep learning model

Our model is inspired from [9]. We decided to get rid of the computer vision part and to focus on a text inputs only: *name*, *username* and *description*. First we train for each input an independent bi-directional long short term memory recurrent neural networks (bi-LSTM RNN) model to predict the gender. For instance, only with the input name, we

Model	F1 - Institutions / People	F1 Men/Women
Wang et al. (2019)	89.90	91.8
Baseline	63.5	57.0
BoW + Logit	78.4	87.3
Ours	78.9	91.5

Table 3. Results of different implementations for 5 languages together

train a bi-LSTM RNN model to predict if the name is more likely to belong to a man, a woman or an organization. Second, we get the three trained models back, discard the final softmax layer of each model, concatenate the last layer of each model together and add a new softmax layer on top of it. We train this architecture in two steps. During the first one, the warm-up step, we freeze all layers but the the final softmax layer and train the model. Then we unfreeze all layers and train all layers together.

We chose bi-LSTM RNN [31] to be aligned with [9] and compare our two ways of constructing our training sets. LSTMs advantage is that they learn long-distance relationships between the inputs.

We trained for each of the inputs a bi-LSTM with character and word embeddings. The character embedding representation enables to represent shared linguistic patterns across languages as suggested in [32]. Besides character embeddings representation with LSTM is pretty efficient to detect morphologically rich language as exposed in [33]. Both embeddings, word and character, will then represent better the meaning of the inputs in a multilingual paradigm.

3. Experimental results

First, we look at the institution detection in 3 on the left. Our best model is not so far from [9]. We reach 87.7% of their results so far with 4.7% of the data volume and without using profile pictures. The deep learning model clearly outperforms the baseline but the classical machine learning model reaches close results. We should increase our training set volume to get the full potential of the deep learning architecture to get results as in [10]: a significant margin between deep learning and classical machine learning methods. Nevertheless, our methodology goes in the good direction as we reach similar results with much lower data.

n-grams	**P**[male]	**P**[female]
cris	0.25	0.75
ina	0.11	0.89
nat	0.39	0.61
isco	0.98	0.02

Table 4. Empiric Probabilities of some *n-grams* to belong to a man or a woman

When we look at the gender differentiation in 3 on the right, we see that our model is almost at the level of [9]. We reach 99.7% of the results with only 4.7% of the data volume they use and without using profile pictures. We also outperform the two other methodologies: the baseline and the classical machine learning by a significant margin.

Indeed the difference between our method including recurrent neural network and the bag of words representation with a logistic regression is higher than 4 points. It confirms results from [10] about significant improvement when detecting gender for deep learning model compared to classical machine learning models.

We also have evidence that our deep learning model learnt well how to differentiate women and men. First, we have computed the empiric probabilities associated with some features such as diminutive names: *cris* has a probability of 75% to be associated with women, as it can also be used by men. We show some examples in Table 4.

4. Conclusions and future work

In our ongoing work we tackle the problem of differentiating people from institution and identifying the gender of real users simultaneously. We propose a soft-labeling based methodology to build our data sets by combining three different methods to tag the collected data from Twitter. In the future, we will add additional soft-labelers, for instance existing classifiers such as [9] or [22] to leverage existing knowledge. The soft-labelers can be adapted to a lot of languages in a relatively small amount of time. We also plan to experiment with calibrated soft-labeling using the methods presented in [34].

We see that our methodology enables to reach good results with a much lower data volume. Besides for gender detection we are almost at the same level than [9] when working on 5 languages and with only text inputs. We also see that the character embeddings paradigm catches morphological variations in a multilingual framework. We will work on adding more data to leverage the deep learning potential, more languages and we will also add the text of the tweet as input as it can carry gender information and also patterns associated to institutions.

We will also provide an ablation study on the different inputs to quantify the importance of each one in the decision. In parallel we will quantify our model gender bias regarding occupations. Our plan is to also provide different labeled test set that anyone can use to benchmark her/his model and to open-source our final model to improve reusability.

References

[1] La fécondité se maintient malgré la Pandémie de Covid-19;. Available from: https://www.insee.fr/fr/statistiques/6024136.

[2] Kemp S. Digital in France: All the statistics you need in 2021 - DataReportal – global digital insights. DataReportal – Global Digital Insights; 2021. Available from: https://datareportal.com/reports/digital-2021-france.

[3] Vasquez-Henriquez P, Graells-Garrido E, Caro D. Tweets on the go: Gender differences in transport perception and its discussion on social media. Sustainability. 2020;12(13):5405. Publisher: Multidisciplinary Digital Publishing Institute.

[4] Bode L, Budak C, Ladd JM, Newport F, Pasek J, Singh LO, et al. Words That Matter: How the News and Social Media Shaped the 2016 Presidential Campaign. Brookings Institution Press; 2020. Available from: https://books.google.es/books?id=JbpkDgAAQBAJ.

[5] Ortega-Sánchez D, Blanch JP, Quintana JI, Cal ESdl, de la Fuente-Anuncibay R. Hate Speech, Emotions, and Gender Identities: A Study of Social Narratives on Twitter with Trainee Teachers. International Journal of Environmental Research and Public Health. 2021;18(8). Available from: https://www.mdpi.com/1660-4601/18/8/4055.

[6] Wood-Doughty Z, Xu P, Liu X, Dredze M. Using noisy self-reports to predict twitter user demographics. arXiv preprint arXiv:200500635. 2020.
[7] Yang YC, Al-Garadi MA, Love JS, Perrone J, Sarker A. Automatic gender detection in Twitter profiles for health-related cohort studies. JAMIA open. 2021;4(2):ooab042. Publisher: Oxford University Press.
[8] Vicente M, Batista F, Carvalho JP. In: Kóczy LT, Medina-Moreno J, Ramírez-Poussa E, editors. Gender Detection of Twitter Users Based on Multiple Information Sources. Cham: Springer International Publishing; 2019. p. 39-54. Available from: https://doi.org/10.1007/978-3-030-01632-6_3.
[9] Wang Z, Hale S, Adelani DI, Grabowicz P, Hartman T, Flöck F, et al. Demographic Inference and Representative Population Estimates from Multilingual Social Media Data. In: The World Wide Web Conference. WWW '19. New York, NY, USA: Association for Computing Machinery; 2019. p. 2056-67. Available from: https://doi.org/10.1145/3308558.3313684.
[10] Liu Y, Singh L, Mneimneh Z. A comparative analysis of classic and deep learning models for inferring gender and age of Twitter users. In: Proceedings of the International Conference on Deep Learning Theory and Applications; 2021. .
[11] Bianchi F, Cutrona V, Hovy D. Twitter-Demographer: A Flow-based Tool to Enrich Twitter Data. arXiv preprint arXiv:220110986. 2022.
[12] Steinert-Threlkeld ZC. Twitter as Data. Elements in Quantitative and Computational Methods for the Social Sciences. Cambridge University Press; 2018.
[13] Ikae C, Savoy J. Gender identification on Twitter. Journal of the Association for Information Science and Technology. 2022;73(1):58-69. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.24541. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.24541.
[14] Rangel F, Rosso P, Koppel M, Stamatatos E, Inches G. Overview of the author profiling task at PAN 2013. In: CLEF Conference on Multilingual and Multimodal Information Access Evaluation. CELCT; 2013. p. 352-65.
[15] Rangel F, Rosso P, Chugur I, Potthast M, Trenkmann M, Stein B, et al. Overview of the 2nd author profiling task at pan 2014. In: CLEF 2014 Evaluation Labs and Workshop Working Notes Papers, Sheffield, UK, 2014; 2014. p. 1-30.
[16] Rangel Pardo FM, Celli F, Rosso P, Potthast M, Stein B, Daelemans W. Overview of the 3rd Author Profiling Task at PAN 2015. In: CLEF 2015 Evaluation Labs and Workshop Working Notes Papers; 2015. p. 1-8.
[17] Rangel F, Rosso P, Verhoeven B, Daelemans W, Potthast M, Stein B. Overview of the 4th author profiling task at PAN 2016: cross-genre evaluations. In: Working Notes Papers of the CLEF 2016 Evaluation Labs. CEUR Workshop Proceedings/Balog, Krisztian [edit.]; et al.; 2016. p. 750-84.
[18] Rangel F, Rosso P, Potthast M, Stein B. Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter. Working notes papers of the CLEF. 2017:1613-0073.
[19] Rangel F, Rosso P, Montes-y Gómez M, Potthast M, Stein B. Overview of the 6th author profiling task at pan 2018: multimodal gender identification in twitter. Working Notes Papers of the CLEF. 2018:1-38.
[20] Rangel F, Rosso P. Overview of the 7th author profiling task at PAN 2019: bots and gender profiling in twitter. In: Working Notes Papers of the CLEF 2019 Evaluation Labs Volume 2380 of CEUR Workshop; 2019. .
[21] Michael J. 40 000 Namen. Anredebestimmung anhand des Vornamens. c't. 2007 Aug;17:182-3. Place: Hannover Publisher: Heise Zeitschriften Verlag. Available from: http://www.heise.de/ct/ftp/07/17/182/.
[22] Knowles R, Carroll J, Dredze M. Demographer: Extremely Simple Name Demographics. In: Proceedings of the First Workshop on NLP and Computational Social Science. Austin, Texas: Association for Computational Linguistics; 2016. p. 108-13. Available from: https://aclanthology.org/W16-5614.
[23] Wood-Doughty Z, Andrews N, Marvin R, Dredze M. Predicting Twitter User Demographics from Names Alone. In: Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media. New Orleans, Louisiana, USA: Association for Computational Linguistics; 2018. p. 105-11. Available from: https://aclanthology.org/W18-1114.
[24] Hu Y, Hu C, Tran T, Kasturi T, Joseph E, Gillingham M. What's in a name? – gender classification of names with character based machine learning models. Data Mining and Knowledge Discovery. 2021 Jul;35(4):1537-63. Available from: https://doi.org/10.1007/s10618-021-00748-6.
[25] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of

the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics; 2019. p. 4171-86. Available from: https://aclanthology.org/N19-1423.

[26] Wood-Doughty Z, Mahajan P, Dredze M. Johns Hopkins or johnny-hopkins: Classifying Individuals versus Organizations on Twitter. In: Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media. New Orleans, Louisiana, USA: Association for Computational Linguistics; 2018. p. 56-61. Available from: https://aclanthology.org/W18-1108.

[27] Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. Transformers: State-of-the-Art Natural Language Processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Online: Association for Computational Linguistics; 2020. p. 38-45. Available from: https://aclanthology.org/2020.emnlp-demos.6.

[28] Kirk H, Jun Y, Iqbal H, Benussi E, Volpin F, Dreyer FA, et al.. Bias Out-of-the-Box: An Empirical Analysis of Intersectional Occupational Biases in Popular Generative Language Models. arXiv; 2021. Available from: https://arxiv.org/abs/2102.04130.

[29] Redmon J, Divvala S, Girshick R, Farhadi A. You Only Look Once: Unified, Real-Time Object Detection. arXiv; 2015. Available from: https://arxiv.org/abs/1506.02640.

[30] Sap M, Park G, Eichstaedt J, Kern M, Stillwell D, Kosinski M, et al. Developing Age and Gender Predictive Lexica over Social Media. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics; 2014. p. 1146-51. Available from: https://aclanthology.org/D14-1121.

[31] Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Comput. 1997 nov;9(8):1735–1780. Available from: https://doi.org/10.1162/neco.1997.9.8.1735.

[32] Chung J, Cho K, Bengio Y. A Character-level Decoder without Explicit Segmentation for Neural Machine Translation. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Berlin, Germany: Association for Computational Linguistics; 2016. p. 1693-703. Available from: https://aclanthology.org/P16-1160.

[33] Faruqui M, Tsvetkov Y, Neubig G, Dyer C. Morphological Inflection Generation Using Character Sequence to Sequence Learning. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, California: Association for Computational Linguistics; 2016. p. 634-43. Available from: https://www.aclweb.org/anthology/N16-1077.

[34] Rizve MN, Duarte K, Rawat YS, Shah M. In Defense of Pseudo-Labeling: An Uncertainty-Aware Pseudo-label Selection Framework for Semi-Supervised Learning. arXiv:210106329 [cs]. 2021 Apr. ArXiv: 2101.06329. Available from: http://arxiv.org/abs/2101.06329.