Can an Online Service Predict Gender? - On the State-of-the-Art in Gender Identification from Texts

Stefan Krüger*
stefan.krueger@upb.de
Heinz Nixdorf Institut / Universität Paderborn
Paderborn, Germany

Ben Hermann†
ben.hermann@upb.de
Heinz Nixdorf Institut / Universität Paderborn
Paderborn, Germany

ABSTRACT

Gender equality initiatives are often faced with a problem: In order to determine whether initiatives are successful the gender of individuals in the target group must be known. As self-identification inherently has the problems that individuals have to respond and results may, therefore, be biased and incomplete, the temptation to use automated gender identification methods is evident. In the scientific literature, multiple sources ranging from the individual's name, their social media choices, biological features (e.g., brain scans or fingerprints), to texts attributed to the individual are used for automated gender identification with varying success. In this paper, we systematically inspect scientific publications for gender prediction based on textual data which are published between January 2017 and January 2019 in order to determine if such approaches may supply viable means to reliably determine an author's gender. However, we find that the best approach in the current state-of-the-art works with an accuracy of only 93.4%. Moreover, we discuss the possible harm that gender identification systems might entail due to their inaccuracy and also given that they are assuming a binary gender model. We conclude that gender identification based on textual data is currently no reliable substitute for self-identification.

CCS CONCEPTS

• **Social and professional topics** → **Gender**.

KEYWORDS

gender equality,gender detection,gender identification

ACM Reference Format:
Stefan Krüger and Ben Hermann. 2019. Can an Online Service Predict Gender? - On the State-of-the-Art in Gender Identification from Texts. In *GE '19: Second Workshop on Gender Equality in Software Engineering, May 27, 2019, Montréal, Canada.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/1122445.1122456

*Both authors contributed equally to this research.
†Both authors contributed equally to this research.

GE '19, May 27, 2019, Montréal, Canada

ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Gender identification (or sometimes gender detection) is an active field of research in machine learning and artificial intelligence. The idea is that based on an artifact (e.g., text, voice recording, image) a trained machine can predict the gender of the person behind that artifact. The results of such a prediction are already used on websites to target advertisement based on gender. This may create perception biases and may proliferate gender stereotypes.

Gender-API.com — a paid service for gender identification based on the name and location of a person — is reporting over 20 million queries per month by March 2018. There are multiple services like this available which have easy-to-use support for multiple languages. These services are based on name databases which have been constructed using automated learning steps as well as manual work. Karimi et al. [12] have inspected these services and databases and found their accuracies ranging from 74% to 91%. This means that even the best approach is incorrectly predicting a person's gender in 9% of all cases. Santamaría and Mihaljevic [21] report that when combining three online services they were able to reach an accuracy of 98%, while, however, leaving roughly 25% of their dataset unclassified. Another problem we (and others [8, 13, 21]) see is the implicit assumption of a binary gender model which discriminates against persons of non-binary genders. Moreover, the given name of a person might not be available. Various collaborative platforms on the Internet (e.g., GitHub) do not require the actual name but rather rely on pseudonyms. In order to make gender identification more robust against these issues, style-based analysis of a user's provided textual artifacts is considered as an alternative [22].

In this paper, we systematically review the scientific literature on gender identification and compare the accuracy of machine-learning approaches for gender recognition relying on textual artifacts. The idea of such machine learning based approaches is to take textual input the user provides over time to query a trained machine for the gender. There are other approaches that aim to determine gender by still images [11], video, or audio data, which we are not inspecting in this paper.

We found that the best approach currently available reaches an accuracy of 93.4%. While the technical effort necessary to reach this accuracy is indeed impressive, the resulting accuracy is still forbiddingly low for many use cases as 7 out of 100 persons will be misgendered with this approach. This number may seem low at first glance, but in cases where the access to services or events is restricted based on the outcome of this classification, these accuracy rates might significantly impede misclassified individuals' social inclusion. For other approaches, accuracies vary between 61% and 80% which is very close to random guesses in a binary model. This will negatively impact the validity of any study conducted on the basis of this prediction.

We also discuss the impact of the approaches and critically reflect on the implications such systems might have on society when adopted in industry. The effect of the proliferation of gender stereotypes when such approaches are used for marketing or business decisions is not negligible. Moreover, the use of a binary gender model automatically discriminates against people identifying as non-binary. Given the state of the art and the ethical considerations presented here, we conclude that such services cannot be used as the basis for scientific studies or statistical surveys either.

2 METHODOLOGY

To identify work in this area, we conducted a light-weight literature search following the approach from Kitchenham et al. [14], however, in a very reduced version. We used the DBLP database as a single source for scientific work and limited our study to works published between January 2017 and January 2019. We have chosen this timeframe to capture the state of the art in the field. We only considered journal, conference, or workshop articles. The keywords used were "gender recognition", "gender detection", "gender identification", and "gender inference". We only considered those publications that used text as input regardless of the source of that text.

In total, our search resulted in 215 papers across the three years of coverage. We filtered papers irrelevant to our analysis, i.e. any systems that do not involve some form of gender recognition based on textual input, by reading title and abstract. During this process, we encountered several approaches that were taking part in gender-identification competitions at PAN 2017, PAN 2018, IberEval 2017, and RusProfiling 2017. We decided to include all other entries from these four competitions into our sample set as well, finally arriving at a total of 59 approaches that require further analysis and classification. We classify each approach in terms of four categories. We investigate which sources for training and test data were used. We further provide the reported accuracy values for each tool. Lastly, we determine the underlying gender model made use of in each of the approaches. For brevity's sake, we abstain from showing all approaches partaking in the four competitions. Instead, we provide the best approach for each individual discipline. We rely on the author's presentation of the experiments they conducted and did not attempt to reproduce them.

3 RESULTS

In Table 1, we list all approaches and our classification of them. In the following summary, we only selectively cite approaches of the four competitions that stand out as most of them are similar with respect to our analysis categories. We refer interested readers to the four respective summary papers of the four challenges [16, 18, 19, 24]. All approaches follow a standard machine-learning approach on textual input. First, they tokenize the input texts. Those using social-media posts as input, also replace special-purpose words such as user names or hashtags with generic markers. The pre-processed input is then processed by machine-learning algorithms for classification, ranging from (deep) neural networks to support-vector machines [24]. The models are finally applied to a test set for evaluation.

We first observe that none of the gender-detection approaches accounts for non-binary genders. A majority of the papers outright states that they view gender as a binary classification problem [5, 18, 19, 24], for others, it is more implicit. Given this limitation, we will, in the following, evaluate these systems with respect to their own goal of classifying people into two gender categories.

Most approaches, including the competitors in the four gender-identification challenges mentioned above, use tweets and other social media posts as input. Their accuracy levels range from single digit to high double digit numbers, with the best ones usually around 80% to 85%. To put this into context, even those approaches misgender around two out of ten people. Only very few approaches surpass 90% of accuracy [5, 17]. The approach by Markov et al. [17] presented at RusProfiling 2017 outperforms all other approaches with a reported accuracy 93.4%, but even it misclassifies seven out of a hundred people.

Bsir and Zrigui [3, 4] report that they are able to perform gender identification on the PAN 2017 corpus of tweets in Arabic with an accuracy of 79.23%, while they only reach 62.1% on a collection of Arabic Facebook posts collected by themselves.

Sanchez-Perez et al. [20] perform a machine-learning-based gender identification on a corpus of Spanish-language news articles containing 5,187 texts from 232 different authors where each author is attributed to at least 10 texts. They report an accuracy of 75.61%.

One notable exception is presented by Company and Wanner [5]. They propose two approaches on long-form texts, one on blog posts, the other on whole books. The latter approach is the second most accurate in our accuracy ranking with 91.78%. However, the approach achieves this result after learning on three books of the corresponding author and still misgenders two out of twenty-five people. For any serious use case in the real world, both the amount of input as well as the accuracy of this approach seem impractical.

Related to this point, we also note that none of the papers motivate their approaches with any practical use case. Some approaches like the one by Company and Wanner [5] also tackle the more comprehensive question of author identification with a similar accuracy rate as their gender-detection. It is unclear to us what purpose a gender identification could serve when the author of a text has been or can be identified as well.

Lastly, we observe that only a few papers discuss the limitations of their proposed approaches. Even in those rare cases, the limitation sections are limited to solely technical limitations (e.g. issues with pre-processing of input [1] or the pros and cons of different learning approaches [20]). The work by Simaki et al. [23] represents a noteworthy exception. They inspect sociolinguistic properties (e.g., use of slang) on a corpus of blog posts. While they found that some properties do have a statistically significant correlation with the author's gender (e.g., period length) other properties (e.g., interrogative forms) do not correlate. The results of their study may help to guide machine learning approaches to select better features. Relatedly, a similarly low number of papers [5, 23] present which features act as significant gender signifiers according to their learning algorithms. The general lack of contextualization of one's work does indicate a lack of both scientific rigor as well as awareness of the cultural context that gender and its forms of expression must be placed into.

Table 1: Classification of Automated Gender Identification Approaches 2017 - 2019

Approach	Training Data	Test Data	Best Reported Accuracy	Gender Model
Team nissim17 at PAN 2017	240,000 Arabic Tweets	160,000 Arabic Tweets	68.31%	Binary
	360,000 English Tweets	240,000 English Tweets	74.3%	Binary
	420,000 Spanish Tweets	280,000 Spanish Tweets	80.4%	Binary
Team miura17 at PAN 2017	120,000 Portuguese Tweets	80,000 Portuguese Tweets	85.75%	Binary
Tellez et al. [25] at PAN 2018	150,000 Arabic Tweets	100,000 Arabic Tweets	81.7%	Binary
Daneshvar and Inkpen [6] at PAN 2018	300,000 English Tweets	190,000 English Tweets	82.21%	Binary
	300,000 Spanish Tweets	220,000 Spanish Tweets	82.0%	Binary
Team deepCybErnet at IberEval 2017	4,319 Catalan Tweets	1,081 Catalan Tweets	48.6%	Binary
González et al. [7] at IberEval 2017	4,319 Spanish Tweets	1,081 Spanish Tweets	68.6%	Binary
Markov et al. [17] at RusProfiling 2017	Tweets from 300 Users	Tweets from 200 Users	68.3%	Binary
	Facebook Posts from 228 users		93.4%	Binary
	Reviews from 776 authors		61.9%	Binary
Bhargava et al. [2] at RusProfiling 2017	Essays from 370 authors		78.4%	Binary
	Texts from 94 authors in Gender Imitation Corpus		66%	Binary
Bsir and Zrigui [3]	240,000 Arabic Tweets	160,000 Arabic Tweets	79.2%	Binary
Bsir and Zrigui [4]	240,000 Arabic Tweets	160,000 Arabic Tweets	79%	Binary
	Facebook Posts from 4,444 users		62.1%	Binary
Sanchez-Perez et al. [20]	5,187 Spanish news articles (using cross-validation)		75.61%	Binary
Company and Wanner [5]	4,284 Journalistic Posts		89.97%	Binary
	48 Novels		91.78%	Binary

4 DISCUSSION

Our analysis of text-based gender-detection systems has revealed serious flaws in currently available approaches. The restrictive binary gender model came as no surprise to us as it is in line with previous research on image-based gender detection systems. Keyes [13] analyzed 58 approaches to automated image-based gender recognition. 55 of which assume gender to be binary. However, due to the prevalence of people of non-binary genders [9, 10, 15], this limitation puts into question the usefulness of these systems even when taking only the technical perspective of developing a functioning gender-detection system. Placed in a broader societal context, the consequences may even be grimmer. Such systems have the potential of furthering the erasure of non-binary people, should they ever be employed in practice. We are not the first to come to these conclusions [8, 21]. Hamidi et al. [8] interviewed 13 trans individuals on image-based gender detection systems, all of whom expressed significant discomfort with such systems.

If gender-detection systems are to be used in practice, one has to consider two important caveats. First, such systems based on a binary model directly imply this model for the considered use case, therefore coming with the aforementioned harms to people of non-binary genders. On top of that, a classification of people into a binary gender model for instance for marketing may proliferate gender stereotypes. Second, due to the lack of accuracy, one has to consider cases of unclassifiable results and misclassifications in any practical setting. If access to systems, services, or events is prohibited due to such a system's classification personal rights might be violated and the affected individuals might be prevented from actively partaking in society. If data from detection systems is used in scientific studies the lack of accuracy might harm the overall validity of the study and the use of such systems should, thus, at least be named in the threats to validity. In our opinion, these circumstances forbid the use of such systems in any use case, including but certainly not limited to the pursuit to gather data to increase gender diversity in specific fields (e.g., computer science).

In the current age of digitalization, we as software engineers are often the drivers behind technical innovation. However, for those innovations to not do more harm than good, researchers need to take into account cultural, societal and political contexts their tools and systems are situated in. The limitations to current gender-detection systems discussed both by us and others [8, 13, 21] demonstrate the shortcomings at least with respect to gender and may suggest a wider lack of awareness for such contexts altogether. These shortcomings must be overcome for the field to drive innovation that does not hurt marginalized groups, but protects and supports them.

5 CONCLUSION

In this paper, we presented a literature survey on the current state of the art in gender identification. We show that the best approach available reaches an accuracy of 93.4%, which might not be sufficient for subsequent statistical data processing. In our discussion, we presented several ethical reservations against such systems and

conclude that these systems are currently not sufficient to replace gender self-identification in studies.

REFERENCES

[1] Francesco Barbieri. 2017. Shared Task on Stance and Gender Detection in Tweets on Catalan Independence - LaSTUS System Description. In *Proceedings of the Second Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2017) co-located with 33th Conference of the Spanish Society for Natural Language Processing (SEPLN 2017), Murcia, Spain, September 19, 2017.* 217–221.

[2] Rupal Bhargava, Gunjan Goel, Anjali Shah, and Yashvardhan Sharma. 2017. Gender Identification in Russian Texts. In *Working notes of FIRE 2017 - Forum for Information Retrieval Evaluation, Bangalore, India, December 8-10, 2017.* 13–16.

[3] Bassem Bsir and Mounir Zrigui. 2018. Bidirectional LSTM for Author Gender Identification. In *Computational Collective Intelligence - 10th International Conference, ICCCI 2018, Bristol, UK, September 5-7, 2018, Proceedings, Part I.* 393–402.

[4] Bassem Bsir and Mounir Zrigui. 2018. Enhancing Deep Learning Gender Identification with Gated Recurrent Units Architecture in Social Text. *Computación y Sistemas* 22, 3 (2018). http://www.cys.cic.ipn.mx/ojs/index.php/CyS/article/view/3036

[5] Juan Soler Company and Leo Wanner. 2018. On the role of syntactic dependencies and discourse relations for author and gender identification. *Pattern Recognition Letters* 105 (2018), 87–95.

[6] Saman Daneshvar and Diana Inkpen. 2018. Gender Identification in Twitter using N-grams and LSA: Notebook for PAN at CLEF 2018. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018.*

[7] José-Ángel González, Ferran Pla, and Lluís-Felip Hurtado. 2017. ELiRF-UPV at IberEval 2017: Stance and Gender Detection in Tweets. In *Proceedings of the Second Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2017) co-located with 33th Conference of the Spanish Society for Natural Language Processing (SEPLN 2017), Murcia, Spain, September 19, 2017.* 193–198.

[8] Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M. Branham. 2018. Gender Recognition or Gender Reductionism?: The Social Implications of Embedded Gender Recognition Systems. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, Montreal, QC, Canada, April 21-26, 2018.* 8.

[9] Jack Harrison, Jaime Grant, and Jody L Herman. 2012. A gender not listed here: Genderqueers, gender rebels, and otherwise in the National Transgender Discrimination Survey. (2012).

[10] Daphna Joel, Ricardo Tarrasch, Zohar Berman, Maya Mukamel, and Effi Ziv. 2014. Queering gender: studying gender identity in 'normative'individuals. *Psychology & Sexuality* 5, 4 (2014), 291–321.

[11] Soon-Gyo Jung, Jisun An, Haewoon Kwak, Joni Salminen, and Bernard Jim Jansen. 2018. Assessing the Accuracy of Four Popular Face Recognition Tools for Inferring Gender, Age, and Race. In *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018.* 624–627.

[12] Fariba Karimi, Claudia Wagner, Florian Lemmerich, Mohsen Jadidi, and Markus Strohmaier. 2016. Inferring Gender from Names on the Web: A Comparative Evaluation of Gender Detection Methods. In *Proceedings of the 25th International Conference Companion on World Wide Web (WWW '16 Companion).* International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 53–54. https://doi.org/10.1145/2872518.2889385

[13] Os Keyes. 2018. The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. *PACMHCI* 2, CSCW (2018), 88:1–88:22.

[14] Barbara Kitchenham and Stuart Charters. 2007. *Guidelines for performing Systematic Literature Reviews in Software Engineering.* Technical Report.

[15] Lisette Kuyper and Ciel Wijsen. 2014. Gender identities and gender dysphoria in the Netherlands. *Archives of sexual behavior* 43, 2 (2014), 377–385.

[16] Tatiana Litvinova, Francisco M. Rangel Pardo, Paolo Rosso, Pavel Seredin, and Olga Litvinova. 2017. Overview of the RUSProfiling PAN at FIRE Track on Cross-genre Gender Identification in Russian. In *Working notes of FIRE 2017 - Forum for Information Retrieval Evaluation, Bangalore, India, December 8-10, 2017.* 1–7.

[17] Ilia Markov, Helena Gómez-Adorno, Grigori Sidorov, and Alexander F. Gelbukh. 2017. The Winning Approach to Cross-Genre Gender Identification in Russian at RUSProfiling 2017. In *Working notes of FIRE 2017 - Forum for Information Retrieval Evaluation, Bangalore, India, December 8-10, 2017.* 20–24.

[18] Francisco M. Rangel Pardo, Paolo Rosso, Manuel Montes-y-Gómez, Martin Potthast, and Benno Stein. 2018. Overview of the 6th Author Profiling Task at PAN 2018: Multimodal Gender Identification in Twitter. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018.*

[19] Martin Potthast, Francisco M. Rangel Pardo, Michael Tschuggnall, Efstathios Stamatatos, Paolo Rosso, and Benno Stein. 2017. Overview of PAN'17 - Author Identification, Author Profiling, and Author Obfuscation. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11-14, 2017, Proceedings.* 275–290.

[20] Miguel A. Sanchez-Perez, Ilia Markov, Helena Gómez-Adorno, and Grigori Sidorov. 2017. Comparison of Character n-grams and Lexical Features on Author, Gender, and Language Variety Identification on the Same Spanish News Corpus. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11-14, 2017, Proceedings.* 145–151.

[21] Lucía Santamaría and Helena Mihaljevic. 2018. Comparison and benchmark of name-to-gender inference services. *PeerJ Computer Science* 4 (2018), e156.

[22] Alexander Serebrenik, Bogdan Vasilescu, and Andrea Capiluppi. 2013. Gender, Representation and Online Participation: A Quantitative Study. *Interacting with Computers* 26, 5 (09 2013), 488–511. https://doi.org/10.1093/iwc/iwt047 arXiv:http://oup.prod.sis.lan/iwc/article-pdf/26/5/488/9644215/iwt047.pdf

[23] Vasiliki Simaki, Christina Aravantinou, Iosif Mporas, Marianna Kondyli, and Vasileios Megalooikonomou. 2017. Sociolinguistic Features for Author Gender Identification: From Qualitative Evidence to Quantitative Analysis. *Journal of Quantitative Linguistics* 24, 1 (2017), 65–84.

[24] Mariona Taulé, Maria Antònia Martí, Francisco M. Rangel Pardo, Paolo Rosso, Cristina Bosco, and Viviana Patti. 2017. Overview of the Task on Stance and Gender Detection in Tweets on Catalan Independence. In *Proceedings of the Second Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2017) co-located with 33th Conference of the Spanish Society for Natural Language Processing (SEPLN 2017), Murcia, Spain, September 19, 2017.* 157–177.

[25] Eric Sadit Tellez, Sabino Miranda-Jiménez, Daniela Moctezuma, Mario Graff, Vladimir Salgado, and José Ortiz-Bejar. 2018. Gender Identification through Multi-modal Tweet Analysis using MicroTC and Bag of Visual Words: Notebook for PAN at CLEF 2018. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018.*