Wiki-Gendersort: Automatic gender detection using first names in Wikipedia

Nicolas Bérubé[1], Gita Ghiasi[+,1], Maxime Sainte-Marie[1], and Vincent Larivière[1,2,3]

[1]Canada Research Chair on the transformations of scholarly communication, Université de Montréal
[2]Centre interuniversitaire de recherche sur la science et la technologie
[3]Observatoire des sciences et des technologies
[+]gita.ghiasi.hafezi@umontreal.ca

ABSTRACT

Gender information is often absent from databases available to scholars, thus hindering the proper problematization, investigation, and answering of various gender-related research questions. Named-based algorithms represent the most simple, yet effective used gender detection methods: such methods proceed by generating first-name-to-gender mapping tables based on user records in a given dataset and then applying such mapping tables ”in reversal” to other databases for completion or validation purposes. The present research aims to develop a gender detection algorithm focusing on the gender detection of eponymous Wikipedia pages and compare its performance to that of other well-known gender detection databases, using the author names indexed in the Web of Science.

1 Introduction

The increasing availability of demographic information contained in both online and offline data sources has allowed for more comprehensive and rigorous analyses of various social phenomena and trends. In the case of gender, data growth has given the scientific community a better glimpse of the scope and extent of gender biases and disparities prevalent in social contexts, phenomena, and groups. However, many database lack the proper information about gender to allow such studies.

Due to this situation and in order to improve both data completeness and accuracy, various gender detection algorithms have been developed, aimed at inferring gender from data already provided. While certain face image processing algorithms were designed[1–3], most literature on this matter however focuses on text-based methods. The domains of Marketing, Humanities, Information Science and Census literature have been particularly prolific in that regard. Some attempts were made at inferring gender from users' browsing patterns and history[4,5]. Many techniques were also proposed to predict social network users' private attributes by exploiting public information within their social network[6–10]. As regards to stylometrics-oriented research, several attempts were made to infer gender from various linguistic features, such as character usage, syntax, functional words, and word frequency[11–16]. Relying on various computational methods, from rule-based to both unsupervised[2,17–22] and supervised[23] algorithms, these different approaches were applied on a wide range of datasets, including emails[21], blogs[13,17,22,24], narratives[20,25,26] and tweets[18,21,23].

The most simple, yet effective and used gender detection methods are however name-based: such methods proceed by generating first-name-to-gender mapping tables based on user records in a given dataset and then applying such mapping tables ”in reversal” to other databases for completion or validation purposes[27]. One database used for mapping table generation is the baby name repository from the the US social Security Administration; this database was notably used to study the relationship between gender and job performance among brokerage firms[28], gender disparities in science[29] and patenting[30–32], as well as to develop a thorough demographic profile of Twitter users[33]. Mapping tables were also generated by crawling Facebook public profile pages from a large and diverse sample of New York City users[34].

First name-based methods are especially effective in coping with accuracy problems due to the distinctive nature of certain sub-populations. Research has indeed shown that the relationship between gender and naming practices changes depending on both country[3] and year of birth[27,35] of individuals. Whereas previously-mentioned machine learning methods would require a different gender detection model for each distinctive sub-population[36], name-based methods can easily cope with these caveats by taking into account relevant geographical and temporal information in generating mapping tables[27]. Alternatively, sub-population biases can also be reduced by complementing name-based approaches with face image processing subroutines[3].

Implementing these enhanced first name-based algorithms is not without its challenges, however. On the one hand, user records must include relevant geographical, temporal, and image data, which is certainly not always the case. Furthermore, this complementary data must be properly extracted and joined with the corresponding first names, a task which can easily become cumbersome, especially with large datasets. In light of these different limitations, public, effective and reliable gender detection algorithms based solely on first names are proving relevant and useful.

But as with any other kind of automatic gender detection, dataset availability represents the most important obstacle to name-based algorithms, as all data must be available beforehand for first-name-to-gender mapping tables to be generated. However, the quantity and scope of publicly available datasets containing all this information, or even only first names and gender, is not as broad as one might think, all the more so if it has to be free.

In order to cope with this accessibility issue, the present research aims to develop a gender detection algorithm based on a well-known, crowd-sourced, publicly available database: Wikipedia. In the next section description of this new Wikipedia-based first name genderization algorithm, called Wiki-Gendersort, after which its performance on the Web of Science (WoS) names is evaluated by comparison to that of other well-known gender detection databases.

2 Methodology

The present algorithm maps genderizes first names rather straightforwardly: using the Wikipedia API for Python, it first extracts and cleans content from Wikipedia pages whose title or specially identified content refer to specific personal names. Following this, it assigns a gender by counting key words contained in these pages based on two successive methods, the second one being used only if the first one does not return any conclusive result.

In the end, one of the following five categories is assigned to each first name : *M* for masculine, *F* for feminine, *UNI* for unisex, *INI* for initials, and *UNK* for unknown. For the present paper however, the calculation of gender probability has been reduced to three possible genders : M for masculine, F for feminine or UNI for unisex. This simplification procedure rests on two hypotheses:

1. The distribution of names with *M* and *F* gender is the same for the set of names with assigned gender than for the whole population (including unknown genders *UNI* or *UNK*).

2. The bias caused by attributing the gender M with 100 % certainty for a name with a low chance of being feminine is counterbalanced by the attribution of the gender F to another name that has a low

chance of being masculine.

Simply stated, the first hypothesis states that the results of our gender assignment algorithm is the same for the subpopulation of ”genderized” names than for the whole population. From a statistical perspective, this hypothesis implies generalizing algorithm output from the sample of conclusive cases to the whole name population. This inference isn’t as impactful as it looks, however, since unisex and unknown names only account for about 11 % of occurrences. In other words, our distribution sample is composed of 89 % of the whole distribution, and is therefore very representative, except for a few specific regional biases. As for the second hypothesis, it simply assumes that two similar error probabilities on each side of any M/F bipartition even each other out. However, error margins may be slightly larger or more numerous on one side, which increases the risk of bias impacting the results. Such impact can be minimized by raising the threshold between unisex (UNI) and definitive gender (M/F) assignment, but setting the bar too high might increase biases caused by the first hypothesis. In this paper, the threshold of 3-to-1 (75/25) has been chosen in order to split the gender probability space halfway between equiprobability (50/50) and both maximal and minimal probability (100/0 or 0/100). Simulations with different thresholds (2-to-1 and 4-to-1) indicate that this choice won’t affect the validity of results by more than a few tenths of a percent.

2.1 First name pre-processing

While not mandatory, this pre-processing phase is recommended to filter poorly-formatted data from the database. Some of the steps are specifically tailored to the Web of Science database, which also contains first names with middle names and initials. Since the first names will be used in Wikipedia searches on page titles, it is important that they do not contain any weird characters. The processing on the first name string is done in eight steps:

1. The first name is converted into its closest ASCII character representation.
2. The first name is split into a sequence of strings with spaces and hyphens acting as delimiters. For example, ”John-Paul” will generate the sequence [”John”, ”Paul”].
3. For all strings in the sequence, if the last character is a period, the second to last character is in uppercase and the third to last character is in lowercase, the two last characters are separated in different strings. For example, ”StL.” will be separated as ”St” and ”L.”.
4. Resulting strings in quotations and parenthesis will be moved to the end of the sequence. Quotation marks and parenthesis are then deleted from the strings.
5. If a string does not contain any letter, it is deleted.
6. If a string ends with a period, the period is deleted.
7. To eliminate initials, the processed name will be the first string of the sequence that follows those two criteria: a) regardless of its length, the string does not contain exactly one letter and b) the string contains at least one vowel. If no strings satisfy both criteria, the first name will be automatically set as an initial (*INI*).
8. The first character of the chosen string is converted to uppercase, and the rest of the characters into lowercase.

9. For each string in the ordered sequence, the gender assignation process described in the next section will be applied. If the assignation is set as male (*M*), female (*F*), the process stops. If the assignation is set as unknown (*UNK*) or unisex (*UNI*), then the assignation is applied to the next string in the sequence.

10. If all the strings in the sequence are set as *UNK* or *UNI*, the name will be identified as *UNI* if at least one string in the sequence was set as *UNI*, or *UNK* is all the strings in the sequence we set as *UNK*.

2.2 Gender assignation

First, a Wikipedia search is conducted with the search function of the API to generate a list of pages, limited to 1,000 results. If a page is a disambiguation page, the first page of the disambiguation page is kept, and the remaining pages are kept in a secondary list. If the all the pages from the main list are used, all secondary lists are then analyzed alternatively, meaning that the first elements of all secondary lists will be analyzed, then the second of all secondary lists, and so on.

Two gender assignation methods are then applied to that list. The second method is used only if the first one does not identify a gender either as masculine (*M*), feminine (*F*) or unisex (*UNI*). If both methods fail, the gender is officially set as unknown (*UNK*).

2.2.1 First method:

1. All pages from the list that starts with the first name followed by a space and an uppercase character are kept. Those pages are normally about a person with the studied first name.

2. The summary text before the first section is then analyzed for all words between spaces, apostrophes, periods, commas and parentheses. If the sum of the number of occurrences of ”he” and ”his” in the summary is equal to or more than three times the sum of ”she” and ”her”, the page is identified as masculine. If the sum of ”she” and ”her” is equal to or more than three times the sum of ”he” and ”his”, the page is identified as feminine. If neither cases happen, the page is skipped.

3. Once the page list is exhausted or 20 pages have been identified and not skipped, the query is stopped.

4. If equal to or more than three quarters of all identified pages are masculine, the first name is set as M. Likewise, if equal to or more than three quarters of all identified pages are feminine, the first name is set as F.

5. If at least one page has been identified but less than three quarters of identified pages have been attributed a specific gender, the first name is set as *UNI*.

6. If no page has been identified, the method is inconclusive, and we move to the second method.

2.2.2 Second method:

As opposed to the previous method analysing the content of pages of which the title contains the name, this method analyses the titles of pages of which the content contains the name.

1. If the sum of the number of occurrences of ”men” and ”male” in all page titles is equal to or more than three times the sum of ”women” and ”female”, the first name is set as M. If the sum of ”women” and ”female” is equal to or more than three times the sum of ”men” and ”male”, the first name is set as F. Those pages are normally about gender-specific sporting events where the first name is in the page’s content.

2. If the number of occurrences is non-zero and neither cases in the previous step happen, the first name is set as *UNI*.

3. If the number of occurrences is zero, the method is inconclusive and the first name is set as *UNK*.

This algorithm was applied to the first names found in the Web of Science database and in all gender detection databases found in the next section. Of the 574,129 first name types found in the database and not identified as initials, 130,645 have been assigned a gender; the remaining name types returned inconclusive results and were thus mapped to the value *UNK*. However, since popular names are more frequent and thus more important to identify than rare ones, distinct names should be weighted by their number of occurrences (tokens) in the database. Table 1 shows the number of names and the proportion of Web of Science database first name tokens involved. Indeed, the 130,645 names that have been identified correspond to 65,81 % of the corpus. Table 2 shows the same variables according to the identified gender.

Table 1. Names and occurrences in each of the gender attribution methods

Method	Number of names	Occurrences (%)
Gender identified (1st method)	65,421	62.41
Gender identified (2nd method)	65,224	3.40
Name identified as initials, gender unknown	N/A	28.22
Name not found, gender unknown	443484	5.98

Table 2. Names and occurrences for each of the identified genders

Gender	Number of names	Occurrences (%)
M	75,486	42.91
F	51,556	17.84
UNI	3,593	5.05
UNK	443,484	5.98
INI	N/A	28.22

The set of names that were qualified as *UNK* represent the possible improvement of our current method of gender identification. However, the fact that this set includes approximately 77 % of names shows that a lot of effort would have to be put in order to improve our identification performance by only 5.98 percentage points.

This possible improvement is comparable to the one that could be done by implementing a probability on gender attribution, which would get rid of the UNI classification and possibly improve our identification performance by a maximum of 5.05 %. However, those improvements are irrelevant if we consider those percentages to be below the threshold for hypothesis one to be true, as discussed previously.

The other 28.22 % of occurrences that were identified as *INI* represent the proportion of the Web of Science database that simply do not include any data about the first name. Therefore, they are considered as unknown, but they represent an intrinsic limitation of any gender attribution method integrally or partly based on first names. Adding this proportion in the percentage threshold for the first hypothesis of our model will bring it from 11.03 % to 39.25 %. However, the set of names that were identified as M or F is still more than 60 % of total occurrences, which is a more than representative sample of our population for hypothesis one to still be credible.

2.3 Limitations of the model

The main limitation of the present model refer to the ever-changing nature of Wikipedia. The current data for this current study was collected in the first half of 2017. Due to its popularity, crowdsourced, and open nature, Wikipedia grows at a rate of approximately 600 new pages each day (https://goo.gl/aenESH). Of course, many of these pages are about specific individuals, which first names all have the potential to alter the algorithm output. But Wikipedia is also affected by the temporal and cultural specificities mentioned in the introduction: due to the international and encyclopedic nature of the website, pages about individuals living in any region of the world and during any given time period can be added from anywhere and anytime, thus greatly affecting the country- and time-specificity of name genders. For example, unisex names that are near the identification threshold of 3-to-1 might change to a definitive gender over time. However, we mentioned previously that this should affect the following performance analysis results by only a few tenths of a percent. Likewise, unknown names might eventually be associated a gender as the number of Wikipedia pages grows. The database is therefore dependant on Wikipedia's reliability, which is why we analyse an important number of pages with multiple methods to have a better stability of our database.

Furthermore, the method can sometimes analyze pages that do not refer to persons, but rather to landmarks or geographic locations named after people. This is why, on top of analyzing multiple pages, only the summary of the pages is considered. Most pages that aren't about a person don't have personal pronouns in that section, unless it's about people who discovered, created or are honoured by the subject of the page. Since those people are often males, our algorithm has a slight bias towards male identification, especially for rare names or locations based on names.

Finally, it should be noted that The Wiki-Gendersort algorithm only consider Wikipedia pages written in English. However, the number English articles are more than 2.5 times the number of German ones, which is the second most popular language on Wikipedia. Therefore, even if most of the pages in other languages will have an English counterpart, the algorithm could have a slight bias against international names, giving them a higher proportion of unknown or male identification for previously mentioned reasons. Also, if names are associated with different genders in different countries, only the amalgam based on the number of pages will be considered. For example, the name ”Jean”, feminine in English but masculine in French, is identified as unisex by our algorithm.

It is our opinion that the impact of these different limitations on algorithmic performance is rather low. To prove this, a performance comparison of Wiki-Gendersort to known databases is presented in the next section.

3 Performance Comparison

Four databases were considered for performance analysis, namely Gender-checker, Gender.c, NamSor and the 2010 U.S. Census databases.

3.1 Comparables

3.1.1 gender.c

The Gender.c package is a free database that also associates a gender to a first name. It contains 46 599 names. However, of those names, only 27 053 are found in the Web of Science database, and they account for 58,8 % of all authorship. It is based on the sexmachine database, which contains a list of 40,000 names. Given a name, Sexmachine makes a guess whether the name is male, mostly male, female, mostly female or unclear. provides detailed information about how popular a first name is in a country and how strongly it is associated with a given gender. Therefore, it enables the disambiguation of names based on the country of origin. The list also provides information for a variety of countries including China and India[33].

3.1.2 gender-checker

The GenderChecker.com database assigns gender to 102,240 first names, which accounts for 64,7 % of the WoS. The main advantage of this database is that it contains names that could be assigned to only one gender (F, M, or Unisex) with a high degree of confidence. This database is based on 2001 and 2011 UK Census data, together with 2011 UN Census data and other online sources.

3.1.3 NamSor

NamSor (`https://goo.gl/CsqEBD`) is a dataset used by Science-Matrix and SheFigures 2015[37]. The NamSor database have either a free or a paid plan, depending on the number and type of queries. It can attribute a gender based on the full name, and therefore needs the last name. Queries were made on the most frequent 1,000,000 full names, without including the ones identified as initials, of the Web of Science database. Those full names contain 77,657 distinct first names, which account for 69,7 % of all authorship. First names that were identified to more than one gender depending on the last name were automatically identified as unisex.

3.1.4 2010 U.S. Census data

Finally, the publicly available data from the 2010 U.S. Census contains the 1,219 masculine names and 4,275 feminine names that compose 90 % of the U.S. population. First names that were identified to more than one gender were given a gender based on the 3-to-1 threshold. For this evaluation, since the less popular masculine name in this database apply to 0.004 % of the population, only feminine names that apply to equal or more than this proportion were kept. Therefore, a remaining 1,996 names were identified in the Web of Science database, and they account for 31,2 % of the authorship.

Each database assigns a gender value (*M*, *F*, *UNI* or *UNK*) to a first name. Note that only the NamSor and Gender.c database have the *UNK* category. For each of these names, our algorithm has been applied to identify its gender between *M*, *F*, *UNI*, *UNK* and *INI*. Table 2 to 5 shows the number of names that were identified by each specific gender of the database and our algorithm. The tables also show the percentage of authorships that are accounted by those names.

Table 3. Percentage of authorship of the Web of Science database and number of names depending of their identification by our Wiki-Gendersort algorithm and the Gender-checker database

Genderchecker	**Wiki-Gendersort**					
	M	**F**	**UNI**	**UNK**	**INI**	**Total**
M	29.57 (13607)	0.18 (1197)	0.26 (433)	0.35 (6239)	2.97 (58)	33.34 (21534)
F	0.53 (2268)	13.63 (11091)	0.58 (882)	0.37 (9489)	0.00 (15)	15.11 (23745)
UNI	8.70 (2533)	2.43 (975)	4.05 (733)	0.06 (600)	1.01 (7)	16.24 (4848)
Total						64.69 (50127)

Table 4. Percentage of authorship of the Web of Science database and number of names depending of their identification by our Wiki-Gendersort algorithm and the Gender.c database

Gender.c	**Wiki-Gendersort**					
	M	**F**	**UNI**	**UNK**	**INI**	**Total**
M	29.82	0.04	0.40	0.14	0.00	30.39
	(10432)	(337)	(236)	(2238)	(1)	(13284)
F	0.41	13.06	1.20	0.14	0.00	14.82
	(700)	(7659)	(612)	(3139)	(0)	(12110)
UNI	4.57	2.16	0.87	0.00	0.00	7.60
	(431)	(247)	(138)	(25)	(0)	(841)
UNK	3.42	0.15	2.31	0.08	0.00	5.96
	(395)	(103)	(177)	(143)	(0)	(818)
Total						58.77
						(27053)

Table 5. Percentage of authorship of the Web of Science database and number of names depending of their identification by our Wiki-Gendersort algorithm and the NamSor database

NamSor	**Wiki-Gendersort**					
	M	**F**	**UNI**	**UNK**	**INI**	**Total**
M	39.70	0.51	2.00	2.03	0.00	44.24
	(17111)	(1514)	(502)	(15887)	(0)	(35014)
F	1.18	16.35	1.69	1,14	0.00	20.35
	(2016)	(9847)	(967)	(11810)	(0)	(24640)
UNI	1.13	0.37	1.27	0.36	0.00	3.13
	(471)	(228)	(115)	(1091)	(0)	(1905)
UNK	0.33	0.27	0.01	1.32	0.00	1.94
	(1630)	(1061)	(87)	(13320)	(0)	(16098)
Total						69.66
						(77657)

Two factors are used to test the reliability of our Wiki-Gendersort algorithm. The first one is the percentage of authorship that are assigned a gender in the database that we can identify with our algorithm. In this case, all *UNI*, *INI* and *UNK* will be considered as unidentified. This percentage doesn't have to be 100 %, but it must be high enough to satisfy the first hypothesis from our introduction, mainly that the results of a study conducted on the set of identified names are the same as on those on the full set of names. The second factor is the proportion of correctly genderized names by our algorithm in the subset of authorship that are attributed a gender in the database. This one should be as close as 100 % as possible, since any deviation is attributed to a false identification. Those factors are presented in Table 6 for all four databases.

The proportion of identified authorship of the NamSor database is 89.39 %. However, even if our algorithm identified lass names than Namsor, it is still enough to be reliable, since this percentage must only be high enough to satisfy our first hypothesis. In addition it is expected for NamSor to identify more names since it uses both first and last names. It should also be noted that our algorithm can also identify some names that NamSor could not. Therefore, out of the 77,657 distinct first names of the most frequent

Table 6. Percentage of authorship of the Web of Science database and number of names depending of their identification by our Wiki-Gendersort algorithm and the US Census database

2010 US Census	**Wiki-Gendersort**					
	M	**F**	**UNI**	**UNK**	**INI**	**Total**
M	20.36 (1110)	0.01 (3)	0.17 (18)	0.00 (10)	0.00 (2)	20.54 (1143)
F	0.02 (8)	9.53 (750)	0.66 (59)	0.00 (6)	0.00 (0)	10.22 (823)
UNI	0.30 (15)	0.02 (3)	0.09 (12)	0.00 (0)	0.00 (0)	0.42 (30)
Total						31.18 (1996)

Table 7. Reliability factors of our Wiki-Gendersort algorithm compared to four known databases.

Algorithm	**Identified authorship (%)**	**Correctly identified authorship (%)**
GenderChecker	90.65	98.39
Gender.c	95.85	98.95
NamSor	89.39	97.07
U.S. Census	97.28	99.91

1,000,000 full names, NamSor could identify 92.73 % of them, and our algorithm could identify 85.90 % of them.

Regarding the proportion of correctly identified names, our algorithm identifies correctly 97.07% of the NamSor names. The remaining 2.93 % are therefore feminine names that were attributed to masculine ones, or vice-versa. It is not obvious to choose which database is better in those cases, especially considering most of those cases relate to Asian names. Indeed, the top 10 names of this particular set are: Hong, Lei, Ji, Fang, Gang, Lu, In, Wan, Fan, Xian. They alone account for 0.47 % of occurrences.

In addition, it is almost impossible to aim for 100 % accuracy on all databases since they sometimes contradict each other. For example, the proportion of correctly identified authorship of GenderChecker on NamSor is 98.4 %, and this factor has been calculated on a set of only 19,738 names as opposed to our algorithm's 30,488 names. Therefore, a factor between 97-99 % on those databases is approximately as high as it can be. The two main limitations are the fact that the algorithm only uses first names, and the reliability of the identification of Asian names.

The GenderChecker and the Gender.c database can both identify around 45-50 % of the Web of Science database. Therefore, a compatibility of more than 98 % on both of them demonstrate the reliability of our Wiki-Gendersort algorithm. Our algorithm can however identify a lot more names and accounts for 60.75 % of the Web of Science database.

4 Conclusion

Our gender identification algorithm based on first names uses public data from Wikipedia pages. It provides a free database of more than 130,000 first names that can be used to attribute a gender on 91.7 % of all first names of the Web of Science since the moment they started collecting them in 2008.

As with enhanced methods, the present algorithm can be refined using geographical, temporal, and image data found on Wikipedia sites could help reduce sub-population bias. The design and testing of an appropriate algorithm would be an interesting matter for future research.

The code and database can be found at https://github.com/nicolasberube/Wiki-Gendersort

References

1. Jain, A., Huang, J. & Fang, S. Gender identification using frontal facial images. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, 4–pp (2005).

2. Baluja, S. & Rowley, H. A. Boosting sex identification performance. *Int. J. computer vision* **71**, 111–119 (2007).

3. Karimi, F., Wagner, C., Lemmerich, F., Jadidi, M. & Strohmaier, M. 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*, 53–54 (International World Wide Web Conferences Steering Committee, 2016).

4. Weiser, E. B. Gender differences in internet use patterns and internet application preferences: A two-sample comparison. *CyberPsychology Behav.* **3**, 167–178 (2000).

5. Hu, J., Zeng, H.-J., Li, H., Niu, C. & Chen, Z. Demographic prediction based on user's browsing behavior. In *Proceedings of the 16th international conference on World Wide Web*, 151–160 (ACM, 2007).

6. Heatherly, R., Kantarcioglu, M. & Thuraisingham, B. Preventing private information inference attacks on social networks. *IEEE Transactions on Knowl. Data Eng.* **25**, 1849–1862 (2013).

7. Lindamood, J., Heatherly, R., Kantarcioglu, M. & Thuraisingham, B. Inferring private information using social network data. In *Proceedings of the 18th international conference on World wide web*, 1145–1146 (ACM, 2009).

8. Mislove, A., Viswanath, B., Gummadi, K. P. & Druschel, P. You are who you know: inferring user profiles in online social networks. In *Proceedings of the third ACM international conference on Web search and data mining*, 251–260 (ACM, 2010).

9. Xu, W., Zhou, X. & Li, L. Inferring privacy information via social relations. In *Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on*, 525–530 (IEEE, 2008).

10. Zheleva, E. & Getoor, L. To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles. In *Proceedings of the 18th international conference on World wide web*, 531–540 (ACM, 2009).

11. Koppel, M., Argamon, S. & Shimoni, A. R. Automatically categorizing written texts by author gender. *Lit. linguistic computing* **17**, 401–412 (2002).

12. Rybicki, J. Vive la difference: Tracing the (authorial) gender signal by multivariate analysis of word frequencies. *Digit. Scholarsh. Humanit.* **31**, 746–761 (2015).

13. Mukherjee, A. & Liu, B. Improving gender classification of blog authors. In *Proceedings of the 2010 conference on Empirical Methods in natural Language Processing*, 207–217 (Association for Computational Linguistics, 2010).

14. Sarawgi, R., Gajulapalli, K. & Choi, Y. Gender attribution: tracing stylometric evidence beyond topic and genre. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, 78–86 (Association for Computational Linguistics, 2011).

15. Peersman, C., Daelemans, W. & Van Vaerenbergh, L. Predicting age and gender in online social networks. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, 37–44 (ACM, 2011).

16. Argamon, S., Koppel, M., Fine, J. & Shimoni, A. R. Gender, genre, and writing style in formal written texts. *TEXT-THE HAGUE THEN AMSTERDAM THEN BERLIN-* **23**, 321–346 (2003).

17. Mikros, G. K. Authorship attribution and gender identification in greek blogs. *Methods Appl. Quant. Linguist.* **21**, 21–32 (2012).

18. Burger, J. D., Henderson, J., Kim, G. & Zarrella, G. Discriminating gender on twitter. In *Proceedings of the conference on empirical methods in natural language processing*, 1301–1309 (Association for Computational Linguistics, 2011).

19. Sboev, A., Litvinova, T., Gudovskikh, D., Rybka, R. & Moloshnikov, I. Machine learning models of text categorization by author gender using topic-independent features. *Procedia Comput. Sci.* **101**, 135–142 (2016).

20. Argamon, S., Goulain, J.-B., Horton, R. & Olsen, M. Vive la différence! text mining gender difference in french literature. *Digit. Humanit. Q.* **3** (2009).

21. Deitrick, W. *et al.* Author gender prediction in an email stream using neural networks. *J. Intell. Learn. Syst. Appl.* **4**, 169 (2012).

22. Bartle, A. & Zheng, J. Gender classification with deep learning (2015).

23. 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*, 37–44 (ACM, 2010).

24. Yan, X. & Yan, L. Gender classification of weblog authors. In *AAAI spring symposium: computational approaches to analyzing weblogs*, 228–230 (Palo Alto, CA, 2006).

25. Weingart, S. & Jorgensen, J. Computational analysis of the body in european fairy tales. *Lit. Linguist. Comput.* **28**, 404–416 (2012).

26. Graliński, F., Jaworski, R., Borchmann, Ł. & Wierzchoń, P. Vive la petite différence! In *International Conference on Text, Speech, and Dialogue*, 54–61 (Springer, 2016).

27. Müller, D., Te, Y.-F. & Jain, P. Improving data quality through high precision gender categorization. URL http://cocoa.ethz.ch/downloads/2018/01/2394_PID5129483.pdf.

28. Green, C., Jegadeesh, N. & Tang, Y. Gender and job performance: Evidence from wall street. *Financial Analysts J.* **65**, 65–78 (2009).

29. West, J. D., Jacquet, J., King, M. M., Correll, S. J. & Bergstrom, C. T. The role of gender in scholarly authorship. *PloS one* **8**, e66212 (2013).

30. Hunt, J., Garant, J., Herman, H. & Munroe, D. Why are women underrepresented amongst patentees? *Res. Policy* **42**, 831–843 (2013).

31. Sugimoto, C. R., Ni, C., West, J. D. & Larivière, V. The academic advantage: Gender disparities in patenting. *PLoS One* **10**, e0128000 (2015).

32. Milli, J., Gault, B., Williams-Barron, E., Xia, J. & Berlan, M. The gender patenting gap. Briefing paper IWPR #C440, Institute for Women's Policy Research, 1200 18th Street, Suite 301, Washington, DC 20036 (2016).

33. Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P. & Rosenquist, J. N. Understanding the demographics of twitter users. *ICWSM* **11**, 25 (2011).

34. Tang, C., Ross, K., Saxena, N. & Chen, R. What's in a name: A study of names, gender inference, and gender behavior in facebook. In *International Conference on Database Systems for Advanced Applications*, 344–356 (Springer, 2011).

35. Blevins, C. & Mullen, L. Jane, john... leslie? a historical method for algorithmic gender prediction. *DHQ: Digit. Humanit. Q.* **9** (2015).

36. 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*, 1136–1145 (2013).

37. EuropeanCommission. She figures 2015. women and science. statistics and indicators (2015). URL https://ec.europa.eu/research/swafs/pdf/pub_gender_equality/she_figures_2015-final.pdf.