UNVEILING GENDER FROM INDONESIAN NAMES USING RANDOM FOREST AND LOGISTIC REGRESSION ALGORITHMS

Musthofa Galih Pradana[1*]; Pujo Hari Saputr [2]; Dyah Listianing Tyas[3]

Informatics, Computer Science Faculty[1]
University Pembangunan Nasional Veteran Jakarta, Jakarta, Indonesia[1]
https://www.upnvj.ac.id/[1]
musthofagalihpradana@upnvj.ac.id[1*]

Informatics Engineering, Engineering Faculty[2]
Sam Ratulangi University, Manado, Indonesia[2]
https://www.unsrat.ac.id/[2]
pujoharisaputro@unsrat.ac.id[2]

Informatics, Science and Technology Faculty[3]
Prisma University, Manado, Indonesia[3]
https://prisma.ac.id/[3]
dyahlistianingtyas@gmail.com[3]

(*) Corresponding Author

Abstract— *Gender detection can be done in many ways, some of these ways by using image identification such as the process of image identification based on faces or image shapes, on the other hand image identification and detection can also be done based on text or written data. The usefulness of gender identification can be used in various aspects of life, ranging from greetings such as ladies and gentlemen, which will certainly make the person concerned feel more appreciated by the accuracy of the pronunciation of the name. This gender identification and detection process can be done by making class predictions on predetermined gender label classes. Of course, each name in various languages has different characteristics in identifying and representing each gender, as well as Indonesian names that have diversity and unique levels of variation. The purpose of this study is to test the results of the algorithm in classification based on class labels. The application of this detection uses two algorithms, namely Random Forest and Logistic Regression. Both of these algorithms can predict classes with perfect accuracy in 6 experimental data, then the results of 526 experimental data resulted in a final accuracy of 0.94 for logistic regression and 0.93 for random forest. The advantage with a thin difference in this case is in the Logistic Regression algorithm.*

Keywords*: detection, gender, logistic regression, random forest, text_classification.*

Intisari—*Pendeteksian gender dapat dilakukan dengan banyak cara, beberapa cara tersebut dengan menggunakan identifikasi citra seperti proses identifikasi citra berdasarkan wajah atau bentuk citra, di sisi lain identifikasi dan deteksi citra dapat pula dilakukan berdasarkan data teks atau tulisan. Kebermanfaatan identifikasi gender ini dapat digunakan dalam berbagai aspek kehidupan, mulai dari pemanggilan sapaan seperti tuan dan nyonya, yang tentunya akan membuat orang bersangkutan dapat merasa lebih dihargai dengan adanya ketepatan penyebutan nama tersebut. Proses identifikasi dan deteksi gender ini dapat dilakukan dengan melakukan prediksi kelas pada kelas label gender yang sudah ditentukan. Tentu setiap nama dalam berbagai bahasa memiliki ciri yang berbeda dalam mengidentifikasikan dan merepresentasikan setiap gender, begitu pula dengan nama berbahasa Indonesia yang memiliki keberagaman dan tingkat variasi yang unik. Tujuan penelitian ini adalah menguji hasil algoritma dalam klasifikasi berdasarkan label kelas. Penerapan deteksi ini menggunakan dua algoritma yakni Random Forest dan Logistic Regression. Kedua algoritma ini dapat memprediksi kelas dengan ketepatan yang sempurna pada 6 data percobaan, kemudian hasil dari* *percobaan 526 data dihasilkan akurasi akhir sebesar 0,94 untuk logistic regression dan 0,93 untuk random forest. Keunggulan dengan nilai selisih tipis pada kasus ini ada pada algoritma Logistic Regression.*

Kata Kunci*: deteksi, jenis kelamin, regresi logistik, hutan acak, klasifikasi teks.*

INTRODUCTION

The process of gender identification and detection is currently widely done. Gender detection can be used in many approaches such as the process of identifying images based on faces or the shape of facial images so that a person's identity is able to be identified and able to represent information that matches a person's identity (Sari, 2022). The process of identification based on images can have implications for detecting a person's identity based on facial images, gender detection can also be done with a text-based data approach (Younis, 2024). The usefulness of this gender identification can be used in various aspects of life, starting from calling greetings such as mr. and madam, which of course will make the person concerned feel more appreciated with the accuracy of the name pronunciation. The direct implementation of this can be seen how the process of sending emails to someone who is just known or a stranger who is not very well known. One example is the job interviewer process where information is often sent through text-based media.

Identification process can be done correctly, then someone who is too well known will feel that it can be appreciated by the accuracy of this gender identification process. Another thing that can be obtained from this text data-based gender identification process is that the diversity of names that are currently diverse can provide an overview of the performance of random forest and logistic regression algorithms in classifying and detecting the diversity of words and names (shaaban, 2022), (Hassan S , 2022), (Lynda, 2023).

These are several relevant studies on the same topic and algorithm. Text-based gender detection research has been conducted based on text from social media twitter or X, the results of this study state that using a word embedding model can significantly improve algorithm performance (Vashisth & Meehan, 2020) Next there is a survey paper that writes about text mining, where one of the papers shows the realm in identifying the relationship between gender, personality, and Twitter addiction (Karami et al., 2020). Kumar's results show that traditional multilabel transformation methods achieve better performance for small amounts of data and remote sequences in terms of samples and labels on text-based gender detection (Kumar et al., 2022). Previous research on machine learning-based gender classification has shown that classification results can provide results such as controlling gender bias in generative models, detecting gender bias in texts, and explaining gender-sensitive language (Dinan et al., 2020). Lexicon-based gender classification research also obtained significant results (Cryan et al., 2020) with a yield of 80% (Bartl & Leavy, 2022).

In another study, it was known that there was a compatibility between two methods, namely the Logistic Regression and MH models, in identifying the function of differential items related to gender on the mathematical ability scale. The scale was designed and tested on a sample of 800 students, consisting of 380 men and 420 women in Jordan. The results of the study show that: (1) the compatibility between the two methods in detecting DIF reaches 80%. (2) Men are superior to women in spatial planning skills and deductive ability, while women are better at numerical ability.

Gender detection is also one aspect of detection on twitter text in the research identification of user profiles based on personal information such as age, personality, gender, education from users' online posts from 6900 Twitter accounts and obtained good and significant results (Karami et al., 2020). Other results showed that gender associations (male-female) with well-studied attributes such as home-work, art-science, math-reading, and good-bad, as well as hundreds of occupational traits and labels, emerged with consistent amounts in children and adults and were well identifiable (Charlesworth et al., 2021).

Word-based detection on gender in Saudi states that Saudi men use different styles that distinguish them from women in terms of politeness (greetings, thank you, apologies, congratulations, encouragement, best wishes etc.), impoliteness (profanity and sarcasm), intensified use, fences, colors, emotions, reasons, emojis, and many others (Alanazi, 2019). Gender detection can also be apart from text such as image detection which can help identify a person's gender and identity (Musthofa Galih Pradana, 2023). Creating author profiles from text documents has been able to be one of the interesting things in the realm of NLP with the results of classical machine learning methods still better than Deep Learning methods for age and gender classification tasks (HaCohen-Kerner, 2022).

Gender detection in a case solved using NLP can be concluded from the importance of the word embedding process in NLP applications, which is able to cause classification of data on social media (Yang et al., 2021) And word embedding is also important to document more comprehensively where the bias exists and may remain hidden, allowing that bias to persist unconsciously across a large corpus of text (Caliskan et al., 2022).

In the realm of algorithms used, the following references show the results of logistic regression and random forest algorithms. The relevant research referred to was from Kanish Shah and the team that carried out the classification by classification in the text. The results show that the logistic regression classifier using the TF-IDF vectorizer function achieves the highest accuracy of 97% on the data set. This algorithm has proven to be the most stable classifier on small data sets (Shah et al., 2020). Random Forest vs Logistic Regression was once written by Kaitlin Kirasich.

The results of each case study of 1000 simulations and the performance of the model consistently show the False Positive Rate Statistics for Random Forest with 100 Trees This is different from Logistic Regression. In all four cases, Logistic Regression and Random Forest achieved different relative classification scores among different simulated logging conditions (Utiarahman, 2024). Comparison of logistic regression, multi-layer perceptron and random forest in the classification of student performance test obtained the result of logistic regression as the best algorithm with an accuracy of 73.9% (Galih Pradana et al., 2023). Weibiao Qiao experimented with his research classification using local wavelet acoustic patterns and Perceptron Multi-Layer neural networks. Based on the results obtained, mWOA classifies sonar data 1.2891 better than GMDH. Score along with other classifications. Overall, the use of MLP-mWOA A study on the classification of passive sonar targets shows this. This algorithm can be used to classify different high-level A-dimensional underwater datasets (Qiao et al., 2021).

MATERIALS AND METHODS

The flow of research carried out can be described as follows :

a. Collecting Data

The data collection process is carried out using a dataset that has been equipped with gender labels. The dataset was obtained from the open dataset available for names in Indonesian Language. Each entry in the dataset includes the names of individuals who are already labeled according to their gender. The existence of this gender label makes it very easy in the classification process, especially in gender detection. With this structured information, classification algorithms can work more effectively and accurately in identifying gender based on the given name.

b. Classification Random Forest

The classification and prediction process was carried out using the Random Forest algorithm to determine the gender class. This algorithm utilizes ensemble learning techniques that combine multiple decision trees to improve accuracy and stability in predicting gender categories from available data.

c. Classification Logistic Regression

The classification and prediction process uses the Logistic Regression algorithm to determine the gender class. This algorithm calculates the probability of a gender category based on input features, allowing for effective and measurable predictions for the analyzed data.

d. Testing

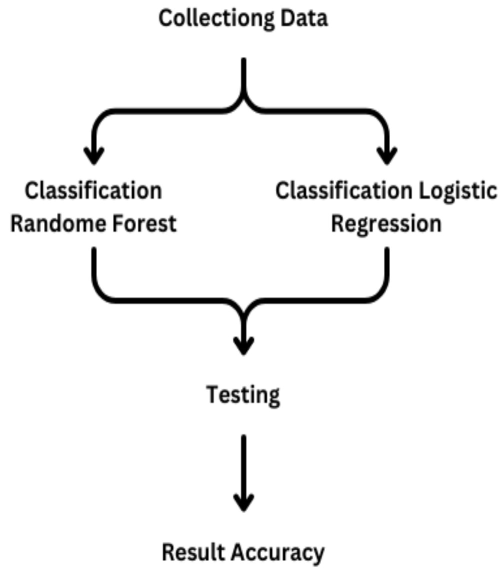
The testing process is carried out by using data that does not yet have a label to determine the prediction of the class. This unlabeled data is tested to predict the appropriate class according to the applied method, allowing for the evaluation of the model's accuracy and effectiveness in identifying the right class label based on available features.

e. Result

The results obtained from the two algorithms are compared to their performance, from the results of the prediction class produced.

The details of this research flow are shown in Source: (Research Results, 2024)

Figure 1.


Collectiong Data
Classification Randome Forest
Classification Logistic Regression
Testing
Result Accuracy

Source: (Research Results, 2024)

Figure 1. Research Stage

From the diata=s research flow, data was used in the study as many as 1316 name data along with labels from gender. The data used in this study is shown in Table 1.

Table 1. Dataset

Num	Name	Gender
1	Erwin Tjahjono	Male
2	Daviandrie Andika Bahroeny	Male
3	Elan Kurniawan	Male
4	Sita	Female
5	Masni Tambunan	Female
6	Marinem	Female
1316	Ngaliman	Male

Source: (Research Results, 2024)

The data is used as a reference in a prediction model carried out to determine the class of unknown gender in Table 2.

Table 2. Prediction Label

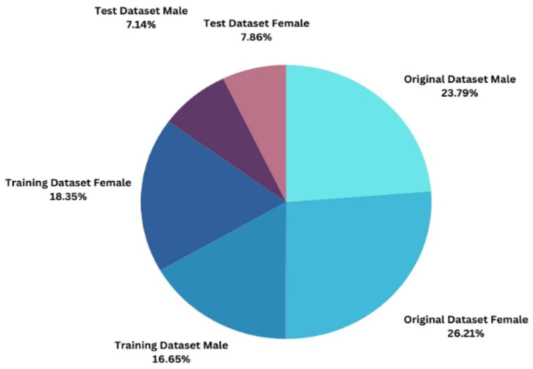
Num	Name	Label
1	Tzalvano	?
2	Andi Wahyudi	?
3	Dimas Raka Bekti	?
4	Cindy Fitria	?
5	Amelia Sari	?
6	Zaskia Widiawati	?

Source: (Research Results, 2024)

The result of the label class prediction regarding this label class will be generated from both algorithms.

RESULTS AND DISCUSSION

The research data used included a total of 1316 samples, which were divided into several categories. The details of the data distribution include the original dataset consisting of male and female samples, as well as the test and training dataset which is also divided by gender. The original dataset included data collected to provide an overview of the distribution of men and women in the study. Meanwhile, a test dataset was used to measure the model's performance on never-before-seen data, with similar details regarding gender. Training datasets, on the other hand, are the data used to train models so that they can learn and make accurate predictions. The results of the visualization of this data distribution can be seen in Source: (Research Results, 2024)
Figure 2., which provides a clear picture of how data is divided and distributed across different categories. Further analysis of these distributions helps in understanding patterns and potential biases in the dataset.



Source: (Research Results, 2024)
Figure 2. Data Distribution

The data distribution is carried out by dividing the data into two parts: training data and test data. After this division, an experimental process is carried out to predict the class of the label using both methods. Predictions were made for the same six names, namely Tzalvano, Andi Wahyudi, Dimas Raka Bekti, Cindy Fitria, Amelia Sari, and Zaskia Widiawati. The prediction results for these names are then compared between two algorithms: Logistic Regression and Random Forest. The results of the prediction using Logistic Regression can be found in Table 3, while the prediction results with Random Forest are displayed on Table 4. This comparison of results helps to evaluate the accuracy and performance of each method in the classification.

Table 3. Prediction Logistic Regression

Num	Name	Prediction	Label
1	Tzalvano	Male	Male
2	Andi Wahyudi	Male	Male
3	Dimas Raka Bekti	Male	Male
4	Cindy Fitria	Female	Female
5	Amelia Sari	Female	Female
6	Zaskia Widiawati	Female	Female

Source: (Research Results, 2024)

Table 4. Prediction Random Forest

Num	Name	Prediction	Label
1	Tzalvano	Male	Male
2	Andi Wahyudi	Male	Male
3	Dimas Raka Bekti	Male	Male
4	Cindy Fitria	Female	Female
5	Amelia Sari	Female	Female
6	Zaskia Widiawati	Female	Female

Source: (Research Results, 2024)

The experiment continued by testing on different names and with a larger quantity of names. The results obtained are shown in Table 5.

Table 5. Data Scenario

Num	Name	Prediction
1	Tzalvano	?
2	Andi Wahyudi	?
3	Dimas Raka Bekti	?
4	Cindy Fitria	?
5	Amelia Sari	?
6	Zaskia Widiawati	?
7	Wahyoedin	?
8	Sukaini	?
9	Rumanah	?
10	Aslam Jusuf	?
30	Rosinta Siregar	?

Source: (Research Results, 2024)

The data were tested into 2 Return methods with the results obtained in Logistic Regression shown in Table 6.

Table 6. Result Logistic Regression

Num	Name	Prediction	Label
1	Tzalvano	Male	Male
2	Andi Wahyudi	Male	Male
3	Dimas Raka Bekti	Male	Male
4	Cindy Fitria	Female	Female
5	Amelia Sari	Female	Female
6	Zaskia Widiawati	Female	Female
7	Wahyoedin	Male	Male
8	Sukaini	Female	Female
9	Rumanah	Female	Female
10	Aslam Jusuf	Male	Male
30	Rosinta Siregar	Female	Female

Source: (Research Results, 2024)

The prediction results using logistic regression with data as much as 30 data, also still show the same results as the detailed data in Table 7.

Table 7. Result Random Forest

Num	Name	Prediction	Label
1	Tzalvano	Male	Male
2	Andi Wahyudi	Male	Male
3	Dimas Raka Bekti	Male	Male
4	Cindy Fitria	Female	Female
5	Amelia Sari	Female	Female
6	Zaskia Widiawati	Female	Female
7	Wahyoedin	Male	Male
8	Sukaini	Female	Female
9	Rumanah	Female	Female
10	Aslam Jusuf	Male	Male
30	Rosinta Siregar	Female	Female

Source: (Research Results, 2024)

The two algorithms used in this analysis are proven to have the same and identical ability in predicting label classes with a very high level of accuracy, according to the actual label class. This shows that both algorithms provide excellent results in the prediction process, making them reliable for applications that require high accuracy.

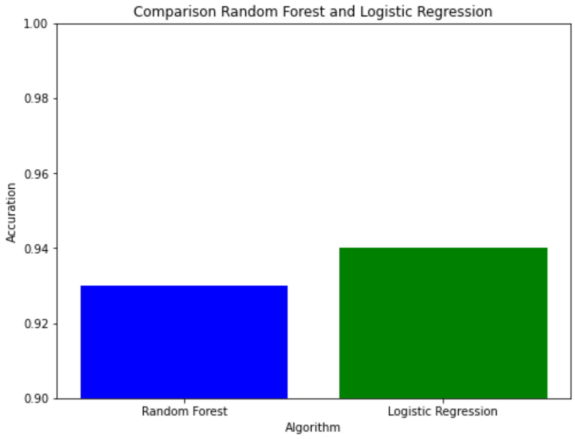
Furthermore, to test the consistency and reliability of the two algorithms in more complex scenarios, experiments were conducted with a larger amount of data. At this stage, the amount of data used increased by 40% from the initial amount of data, bringing the total to around 526 data. This increase in the amount of data aims to observe how the algorithm adapts to greater data variations and whether prediction performance remains stable or undergoes significant changes.

The results of the prediction process carried out using a larger amount of data have been presented in Source: (Research Results, 2024)

Figure 3. At this stage, the data used increases significantly, providing an opportunity to assess how both algorithms are adapting to larger volumes of data. Source: (Research Results, 2024)

Figure 3 menyajikan hasil prediksi yang dilakukan dengan the same algorithm but on a larger dataset, allowing for an in-depth analysis of the effectiveness of both algorithms under those conditions. Analysis of Source: (Research Results, 2024)

Figure 3 will provide additional insight into each algorithm's ability to handle larger amounts of data, and help identify potential performance differences that may arise as the amount of data increases. It is important to evaluate the stability and accuracy of algorithms in more complex and diverse situations, so that they can provide valuable information for the development of better prediction models in the future.

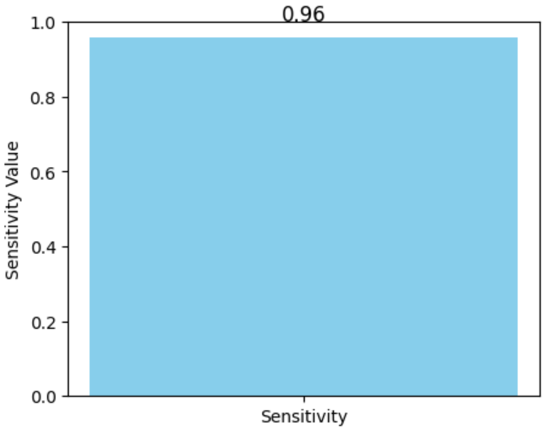


Source: (Research Results, 2024)

Figure 3. Result

In detail, the results of the two algorithms show that the random forest is at 0.93 and the logistic regression is at 0.94. There is a slight difference in the accuracy results produced by the two algorithms with the advantage in logistic regression. The classification carried out is in the context of binary classification, so this logistic regression model tends to have superior accuracy, because indeed this algorithm is designed to calculate the probability of an event occurring based on the value of the feature. This result is reinforced by the sensitivity value of the logistic regression algorithm which reaches a value of 0.96, meaning that the model is able to provide information about how well the model captures all the truly positive data from the total positive data. The result show in Source: (Research Results, 2024)

Figure 4.



Source: (Research Results, 2024)

Figure 4. Sensitivity Logistic Regression

The results of this study can be used as an illustration for the process of gender identification based on name, which implications can be applied in the natural language processing process, for example automatic greetings in email marketing or content marketing.

CONCLUSION

Based on research that has been conducted on gender identification and prediction based on Indonesian names, it was found that there are differences in results between the Random Forest and Logistic Regression methods. This study shows that the Logistic Regression method gives slightly better results compared to Random Forest. Specifically, the accuracy obtained from the Logistic Regression model reached 0.94, while the accuracy of the Random Forest model was 0.93. This small difference in accuracy suggests that although both methods have good performance, Logistic Regression is slightly superior in the context of gender prediction of Indonesian names. The results are expected to be used as an illustration for the process of gender identification based on name, which implications can be applied in the natural language processing process, for example automatically creating greetings in email marketing or content marketing.

The suggestion that can be given based on the results of this study is to increase the variety of data in the training dataset. By increasing the variety of names used in the training data, the prediction process can be improved and updated to be more accurate and representative of the diversity of names in Indonesia. Expanding the training data with more varied names will help in reducing bias and improving the model's performance in identifying gender more effectively and precisely.

REFERENCE

Alanazi, S. A. (2019). Toward Identifying Features for Automatic Gender Detection: A Corpus Creation and Analysis. *IEEE Access*, *7*, 111931–111943. https://doi.org/10.1109/ACCESS.2019.2932026

Bartl, M., & Leavy, S. (2022). Inferring Gender: A Scalable Methodology for Gender Detection with Online Lexical Databases. *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, 47–58. https://doi.org/10.18653/v1/2022.ltedi-1.7

Caliskan, A., Ajay, P. P., Charlesworth, T., Wolfe, R., & Banaji, M. R. (2022). Gender Bias in Word Embeddings. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 156–170. https://doi.org/10.1145/3514094.3534162

Charlesworth, T. E. S., Yang, V., Mann, T. C., Kurdi, B., & Banaji, M. R. (2021). Gender Stereotypes in Natural Language: Word Embeddings Show Robust Consistency Across Child and Adult Language Corpora of More Than 65 Million Words. *Psychological Science*, *32*(2), 218–240. https://doi.org/10.1177/0956797620963619

Cryan, J., Tang, S., Zhang, X., Metzger, M., Zheng, H., & Zhao, B. Y. (2020). Detecting Gender Stereotypes: Lexicon vs. Supervised Learning Methods. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–11. https://doi.org/10.1145/3313831.3376488

Dinan, E., Fan, A., Wu, L., Weston, J., Kiela, D., & Williams, A. (2020). *Multi-Dimensional Gender Bias Classification*. http://arxiv.org/abs/2005.00614

Galih Pradana, M., Palilingan, K., Vanli Akay, Y., Puspasari Wijaya, D., & Hari Saputro, P.

(2023). *Comparison of Multi Layer Perceptron, Random Forest & Logistic Regression on Students Performance Test*. 462–466. https://doi.org/10.1109/icimcis56303.2022.10017501

HaCohen-Kerner, Y. (2022). Survey on profiling age and gender of text authors. *Expert Systems with Applications*, *199*, 117140. https://doi.org/10.1016/j.eswa.2022.117140

Karami, A., Lundy, M., Webb, F., & Dwivedi, Y. K. (2020). Twitter and Research: A Systematic Literature Review Through Text Mining. *IEEE Access*, *8*, 67698–67717. https://doi.org/10.1109/ACCESS.2020.2983656

Kumar, J. A., Trueman, T. E., & Cambria, E. (2022). Gender-based multi-aspect sentiment detection using multilabel learning. *Information Sciences*, *606*, 453–468. https://doi.org/10.1016/j.ins.2022.05.057

Musthofa Galih Pradana, H. K. (2023). Analisis Performa Algoritma Convolutional Neural Networks Menggunakan Arsitektur Lenet Dan Vgg16. *Indonesian Journal of Business Intelligence (IJUBI)*, *6*(2), 54–60.

Qiao, W., Khishe, M., & Ravakhah, S. (2021). Underwater targets classification using local wavelet acoustic pattern and Multi-Layer Perceptron neural network optimized by modified Whale Optimization Algorithm. *Ocean Engineering*, *219*(June 2020), 108415. https://doi.org/10.1016/j.oceaneng.2020.108415

Sari, Y. (2022). *Ekstraksi Fitur dan Aplikasinya pada Citra 2D*. Perahu Litera.

Shah, K., Patel, H., Sanghvi, D., & Shah, M. (2020). A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification. *Augmented Human Research*, *5*(1). https://doi.org/10.1007/s41133-020-00032-0

Yang, Y.-C., Al-Garadi, M. A., Love, J. S., Perrone, J., & Sarker, A. (2021). Automatic gender detection in Twitter profiles for health-related cohort studies. *JAMIA Open*, *4*(2). https://doi.org/10.1093/jamiaopen/ooab042