Damegender: Writing and Comparing Gender Detection Tools

David Arroyo Menéndez, Jesús M. González Barahona and Gregorio Robles

June 29, 2020

About the thesis

- Thesis Student: David Arroyo Menéndez
- Title: Writing and Comparing Gender Detection Tools from a Name
- Thesis Director: Jesús González Barahona
- Objectives:

To compare APIs to detect gender from a name
To apply machine learning to classify nicknames, diminutives or new names as male or female
To understand the impact about open data in statistics about names and gender versus datasets created by companies or free software communities.
To apply this tools to the scientific communities or communities based on Internet.

Damegender in few words (I)

Damegender is:

- A gender detection tool from the name
- Open datasets from official statistics
- Free Software

Damegender can be applied for:

- Gender classification in datasets (Software Repositories, Science, Wikipedia, Mailing Lists, . . .)

Damegender in few words (II)

The main innovations with similar propietary tools are:

- Detect gender in dimininutives and nicknames using ML
- Integration with Perceval to be applied in Internet Communities

The social impact is about:

- Due to the importance about sex variable in sociology
- It's an important problem in Natural Language Processing
- To reduce the gender is an objective in United Nations. You need data and calculus to reach it.!

Download source and article to a make a good tracing

- git clone `https://github.com/davidam/damegender.git`

Introduction (I): Gender detection tools in the bibliography

- The most used method to infer the gender of participants looking for their name.
- To infer the gender from faces in images [RPC17],
- To infer gender from hand-written annotations [LSB11],
- To infer gender from speeches [KAS02].

Introduction (II): Research

- we evaluate the quality (and accessibility, including price) of different existing solutions;
- we discuss their limitations; and
- we investigate what happens with those names not included in official statistics, for example, nicknames or diminutives

Introduction (III): Contributions

- an evaluation of the quality of different solutions applying well-known metrics;
- a tool, called damegender, guessing gender from a name giving support to Spanish and English from the open data census provides by the states built to understand current technologies in detail; this tool has been compared with APIs using an international dataset with good results; and
- a machine learning solution to strings not found in the census dataset to approach the problem with nicknames and diminutives;

Damegender (I): Technologies

- Natural Language Toolkit (NLTK) for natural language processing [LB02]
- Scikit for machine learning [PVG+11],
- Numpy nor numerical computation [VDWCV11], and
- Matplotlib to visualize results [Hun07].
- At its current point it is linked to Perceval [DCRGB18]

Datasets (I): Ways to build a good dataset about names

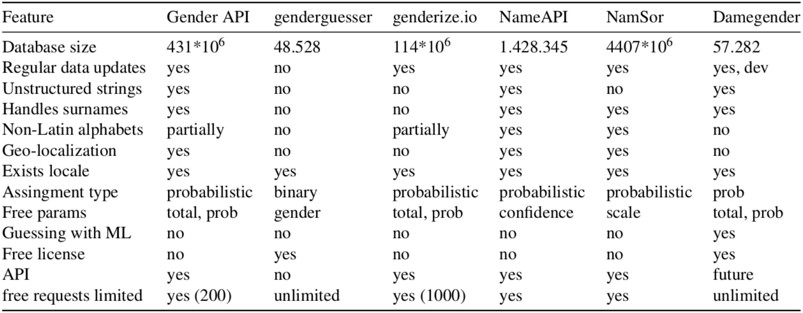
- A census published with a free license (open census way),
- A dataset released with a free license in a free software package (free software way),
- A dataset retrieved from commercial APIs (commercial API way), and
- A dataset which is the result of an investigation and that has been released publicly (scientific way).

Datasets (II): Datasets about names for official statistics in Damegender

- North America: USA and Canada
- South America: Uruguay
- Europe: Ireland, United Kingdom, Spain, Portugal, Iceland, Finland
- Oceania: Australia, New Zealand

Note: the results in this paper has been reached with Spain, Uruguay, USA and United Kingdom official statistics.

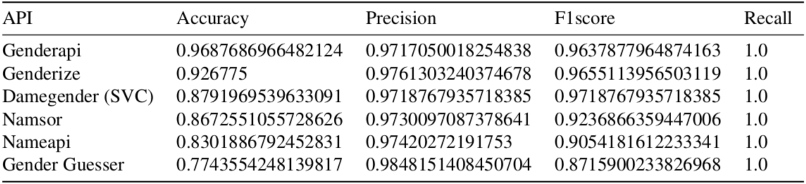
Comparison of the different features that name-to-gender inference services



Feature	Gender API	genderguesser	genderize.io	NameAPI	NamSor	Damegender
Database size	$431*10^6$	48.528	$114*10^6$	1.428.345	$4407*10^6$	57.282
Regular data updates	yes	no	yes	yes	yes	yes, dev
Unstructured strings	yes	no	no	yes	no	yes
Handles surnames	yes	no	no	yes	yes	yes
Non-Latin alphabets	partially	no	partially	yes	yes	no
Geo-localization	yes	no	no	yes	yes	no
Exists locale	yes	yes	yes	yes	yes	yes
Assingment type	probabilistic	binary	probabilistic	probabilistic	probabilistic	prob
Free params	total, prob	gender	total, prob	confidence	scale	total, prob
Guessing with ML	no	no	no	no	no	yes
Free license	no	yes	no	no	no	yes
API	yes	no	yes	yes	yes	future
free requests limited	yes (200)	unlimited	yes (1000)	yes	yes	unlimited

Table 1. Features and gender detection tools by name

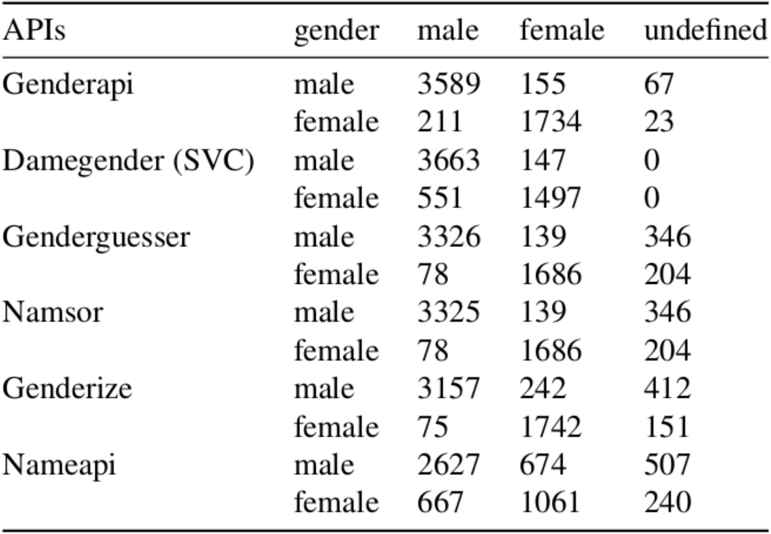
Comparison of measures of the quality of the results for the tools under study (I)



API	Accuracy	Precision	F1score	Recall
Genderapi	0.9687686966482124	0.9717050018254838	0.9637877964874163	1.0
Genderize	0.926775	0.9761303240374678	0.9655113956503119	1.0
Damegender (SVC)	0.8791969539633091	0.9718767935718385	0.9718767935718385	1.0
Namsor	0.8672551055728626	0.9730097087378641	0.9236866359447006	1.0
Nameapi	0.8301886792452831	0.97420272191753	0.9054181612233341	1.0
Gender Guesser	0.7743554248139817	0.9848151408450704	0.8715900233826968	1.0

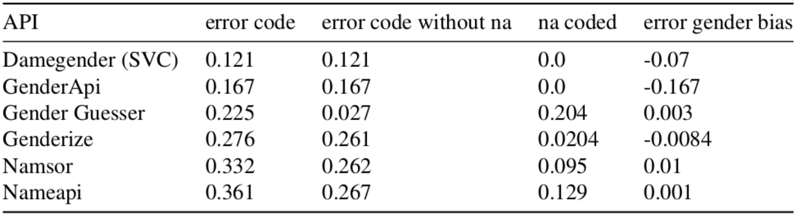
Table 2. Different accuracies measures

Comparison of measures of the quality of the results for the tools under study (II)



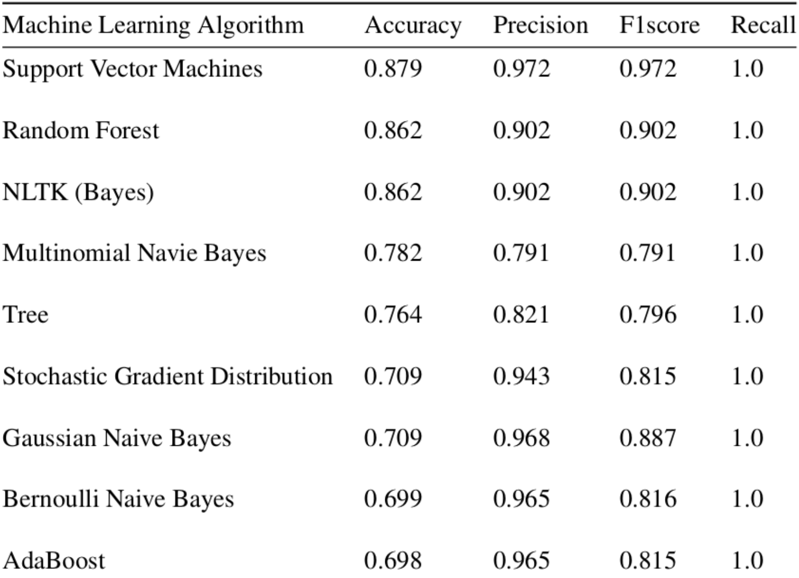
APIs	gender	male	female	undefined
Genderapi	male	3589	155	67
	female	211	1734	23
Damegender (SVC)	male	3663	147	0
	female	551	1497	0
Genderguesser	male	3326	139	346
	female	78	1686	204
Namsor	male	3325	139	346
	female	78	1686	204
Genderize	male	3157	242	412
	female	75	1742	151
Nameapi	male	2627	674	507
	female	667	1061	240

Comparison of measures of the quality of the results for the tools under study (III)



API	error code	error code without na	na coded	error gender bias
Damegender (SVC)	0.121	0.121	0.0	-0.07
GenderApi	0.167	0.167	0.0	-0.167
Gender Guesser	0.225	0.027	0.204	0.003
Genderize	0.276	0.261	0.0204	-0.0084
Namsor	0.332	0.262	0.095	0.01
Nameapi	0.361	0.267	0.129	0.001

Comparison of machine learning algorithms and accuracies



Machine Learning Algorithm	Accuracy	Precision	F1score	Recall
Support Vector Machines	0.879	0.972	0.972	1.0
Random Forest	0.862	0.902	0.902	1.0
NLTK (Bayes)	0.862	0.902	0.902	1.0
Multinomial Navie Bayes	0.782	0.791	0.791	1.0
Tree	0.764	0.821	0.796	1.0
Stochastic Gradient Distribution	0.709	0.943	0.815	1.0
Gaussian Naive Bayes	0.709	0.968	0.887	1.0
Bernoulli Naive Bayes	0.699	0.965	0.816	1.0
AdaBoost	0.698	0.965	0.815	1.0

Conclusions

The market of gender detection tools is dominated by companies based on payment services through APIs. This market could be changed thanks to free software tools and open data due to give more explicative results for the user. Although the machine learning techniques is not new in this field, it's an incentive for researchers in computer science create free software tools.

These advances in computer science could be giving support to study the gender gap in repositories and mailing lists. So, the application of Damegender in real cases is the next step in this research.