Discussion Paper

Profiling of Twitter users: a big data selectivity study

The views expressed in this paper are those of the author(s) and do not necessarily reflect the policies of Statistics Netherlands

2016 | 06

Piet J.H. Daas
Joep Burger
Quan Le
Olav ten Bosch
Marco J.H. Puts

Content

Summary

Big data may contain traces of human or economic activity that could potentially be used for official statistics. On the other hand, big data does not contain a random sample of the target population, which may result in biased estimates. In sample surveys, auxiliary information known for the target population is traditionally used to correct for selective non-response. A similar approach could be applied to big data, if auxiliary information can be extracted and linked to the units in the big data source. In this paper, we explore different ways of extracting auxiliary information, both from the big data source itself and from linking it with other sources of information. We apply this profiling method to a dataset of Dutch Twitter users. We show to what extent gender can be extracted from the user’s first name, short biography, tweet writing style and profile picture. We also show to what extent Twitter accounts can be matched with LinkedIn accounts, from which additional characteristics can be extracted using a web scraping robot. We discuss the potential and implications of profiling big data sources for official statistics.

1. Introduction

In our modern world more and more data are being created and remain stored. These kinds of data, generally referred to as big data, are very interesting sources of information. They, for instance, may reflect traces of human or economic activity and could possibly be used for official statistics (Glasson et al. 2013). However, extracting information from big data for such purposes is challenging for a number of reasons. First, not all data are relevant for the research question at hand, which requires one to find the signal in the noise (Silver 2012). Second, most big data available are composed of events (Daas et al. 2015) and usually provide very little or no information on the unit that generated the data. Third, if information is available on the creator of the data it may not be easily linked to a specific person or company. Fourth, not all units in the target population that the researcher envisaged may be included in big data and the ones that are included are not a random sample from the target population. All in all, these issues make it challenging, to say the least, to use big data for the creation of official statistics. In this paper we will mainly focus on the fourth challenge, i.e. how to assess the selectivity of a big data source.

Let’s illustrate the research question at hand with an example: social media. Many people in the Netherlands are active on social media: 70% of the population posted messages according to a European study (Eurostat 2013). Compared with a probability sample this is an extremely high coverage rate. In contrast to a probability sample, however, we do not know to what extent these social media accounts represent our target population: for social statistics the target population are the persons included in the population register of the Netherlands, while for business statistics these are the companies in the Dutch statistical business register. Quite a number of social media accounts actually reflect the activity of companies (even though they are created by humans). Hypothetically, cyber savvy and extravert people are more likely to be active on social media than computer novices and introverts. In addition, not all activity is publicly available as some social media messages are private only.

These are not uncommon issues as selective non-response in sample surveys also causes a deviation from representativeness. Without correction for selectivity, estimates will be biased. A common method used to assess selectivity in sample surveys is by comparing the distribution of relevant auxiliary variables in the data source with their known distribution in the target population. In principle, the same approach could be applied to big data, although in practice this is not trivial (Buelens et al. 2014, 2015). In an ideal world, units are linked to a population register containing auxiliary variables. Our experiences on studying big data sources have revealed that many units hardly provide any information that could be used to deterministically link them to a population register. In a more realistic big data context, auxiliary information will have to be derived from the big data source itself or by using an additional source of information.

What kinds of characteristics are needed? In many surveys a similar set of characteristics is used that correlate with target variables. In social statistics commonly used variables are: gender, age, income, education, origin, degree of urbanization, and household composition. For companies often used characteristics are: number of employees (size class), turnover, type of economic activity, and legal form. The key question here is how these characteristics should be obtained. This was the starting point of our study.

1.1 Aim of the study

Obtaining auxiliary information from units in big data sources is challenging. In our opinion a method called 'profiling' is an interesting option. This term refers to an approach from the field of information science. In this approach, large amounts of data are analyzed with the aim of discovering patterns to discern groups of similar units (Hildebrandt and Gutwirth 2013). This can be done by i) studying the big data source for 'clues' or by ii) combining the big data source with another source that contains these characteristics. Both approaches were studied here. In this paper, we describe the results of these studies in which social media, Twitter and LinkedIn, were investigated. Advantage of social media is that, from an experimental point of view, a lot of data is publicly available and each unit in the population has a unique identifier: a user id. This in contrast to many other big data sources. Their data may be owned by private companies or they may have computers or other electronic devices as units (Glasson et al. 2013).

For these studies only data available on public Twitter and LinkedIn accounts are used, meaning that anyone with a PC, a browser and an internet connection can access the data studied. In an earlier study performed at Statistics Netherlands in cooperation with Erasmus University (Daas et al. 2012) we obtained a list of 330.000 Twitter usernames that were—at that point in time and according to the location information on their user profile—all identified as Dutch Twitter users. This list is the starting point for the studies described below. From this list a random sample was selected and studied. In Section 2 the results of various ways of 'profiling' the gender of these users from the available Twitter data are described. In Section 3 we describe the results of the study in which we combine the sample of Twitter users with their accompanying publically available LinkedIn profiles. In Section 4, the findings are discussed and conclusions are drawn.

2. Auxiliary information from Twitter itself

A random selection of 1000 Twitter accounts was obtained from a previously collected list of 330.000 usernames (Daas et al. 2012). Since this study was performed several years ago, we first checked if these accounts still existed online. This was done by determining if the webpage of each Twitter account could be accessed. This resulted in a set of 844 still existing accounts that were used in the remainder of this study. These accounts all had a screen name, username and a unique id. Of these accounts 583 provided a short biography, 473 had ever created messages (‘tweets’), and 804 had a non-default picture. A person that does not upload a picture on its profile will have an egg as a default picture. The importance of this will become clear later on in this document.

On Twitter both persons and companies are active. They both represent different populations from an official statistical perspective. For this study it was decided not to include the automatic differentiation between these types of accounts. We manually checked if the account belonged to a person and focused on the profiles of persons. Since the main goal of the first part of our study was to determine the gender of a Twitter account, we created a test set by manually annotating the gender of each person. During this manual checking all available information on the person’s Twitter page and any other webpages referred to from this page, was used to determine if the username belonged to a male, female or other. Examples of the latter are accounts of companies, organizations, animals and bots. The manual classification revealed the composition of the 844 user accounts shown in Table 1.

Table 1. Results of the manual gender classification of 844 selected Twitter accounts.

Gender	Number of accounts	Share (%)
Male	409	49
Female	282	33
Other	153	18

From Table 1 it is clear that 691 persons are included in the person’s dataset. During the manual classification, two accounts that clearly belonged to a person were excluded because the gender of those persons could not be determined with certainty. We decided to assign these accounts to the Other category. The 691 persons accounts are used in the gender determination studies described below.

The manual checking revealed that gender-related information was provided by the Twitter user’s first name, the profile biography, the profile picture, and any information available on webpages that were referred to. Examples of the latter are a personal webpage or an associated LinkedIn account. This showed that Twitter data itself does indeed provide information on someone’s gender. Three of these options were selected for automatic classification, i.e. the user’s first name, the profile biography and the profile picture. Additionally screening of associated pages was initially excluded because in this part of the study we only wanted to study the information provided by Twitter; Section 3 discusses the use of LinkedIn pages. A fourth alternative, not used in the manual classification, was additionally included and studied in cooperation with Dong Nguyen from the University of Twente. Dong and her co-authors developed a classifier that is able to deduce someone's gender (and age) from the writing style in their Dutch tweets (Nguyen et al. 2013). A demo of this work can be found on tweetgenie.nl. We used the API of this service to determine the gender of the Twitter accounts selected. In the next section the results of each of the four approaches will be shown and discussed.

We treat determining the gender of a person as a binary classification (someone is either male or female), i.e. we are ignoring gender neutrality here. However, if someone's gender cannot be determined by the classifier, a third category is introduced: unknown. This makes it somewhat more complicated to calculate, for instance, the accuracy and sensitivity of a classifier, since the calculation of the number of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) cannot be calculated straightforward. To enable this we made the following adjustments as shown in Table 2:

Table 2. Overview of different classifications discerned for the scores for gender.

		Classified		
		Male (1)	**Female (0)**	**Unknown (−1)**
True	Male	MM	MF	MU
	Female	FM	FF	FU

MM: The number of males (correctly) classified as male
MF: The number of males (incorrectly) classified as female
MU: The number of males (incorrectly) classified as unknown
FM: The number of females (incorrectly) classified as male
FF: The number of females (correctly) classified as female
FU: The number of females (incorrectly) classified as unknown

From this we can derive the following performance measures (see e.g. Shaikh 2011):

Accuracy:

$$a = \mathbb{P}\{\text{classified correctly}\} = (MM + FF)/(MM + MF + MU + FM + FF + FU)$$

Sensitivity:

$$s^{\mathrm{M}} = \mathbb{P}\{\text{classified male}|\text{true male}\} = MM/(MM + MF + MU)$$

$$s^{\mathrm{F}} = \mathbb{P}\{\text{classified female}|\text{true female}\} = FF/(FM + FF + FU)$$

Precision:

$$p^{\mathrm{M}} = \mathbb{P}\{\text{classified correctly}|\text{classified male}\} = MM/(MM + FM)$$

$$p^{\mathrm{F}} = \mathbb{P}\{\text{classified correctly}|\text{classified female}\} = FF/(MF + FF)$$

Harmonic mean of sensitivity and precision:

$$F_1^{\mathrm{M}} = 2s^{\mathrm{M}}p^{\mathrm{M}}/(s^{\mathrm{M}} + p^{\mathrm{M}})$$

$$F_1^{\mathrm{F}} = 2s^{\mathrm{F}}p^{\mathrm{F}}/(s^{\mathrm{F}} + p^{\mathrm{F}})$$

Diagnostic Odds Ratio:

$$DOR = \frac{s^{\mathrm{M}}s^{\mathrm{F}}}{(1-s^{\mathrm{M}})(1-s^{\mathrm{F}})} = \frac{MM/(MF+MU)}{(FM+FU)/FF}$$

Like accuracy, the DOR is a single performance measure but has the advantage over accuracy that it is independent of the sex ratio in the sample. The DOR ranges between 0 and infinity, where 1 is the null hypothesis corresponding to random guessing. Usually, the (natural) logarithm is taken to arrive at a symmetrical scale ranging from minus infinity to infinity where 0 is the null hypothesis, i.e. random guessing.

A consequence of the adjustments made to the performance measures is that wrongly classified gender and unclassified gender both affect the outcome of the measures. Approaches that result in many unclassified persons will have rather low overall performance although they may, for instance, be very precise (see e.g. Section 2.2). We therefore list all performance measures for each approach and discuss their interpretation. As a benchmark, we show the performance measures of randomly assigning gender to persons and of assigning all persons as either male, female or unknown (Table 3). Because of the absence of some groups not all performance measures can be calculated.

Table 3. Performance measures for randomly assigning gender and assigning all persons as male, female or unknown.

Measure	Perspective	Random	All male	All female	All unknown
Accuracy (%)		50	59.2	40.8	0
Sensitivity (%)	Male	50	100	0	0
	Female	50	0	100	0
Precision (%)	Male	59	59.2	-	-
	Female	41	-	40.8	-
F1 (%)	Male	54	74.4	-	-
	Female	46	-	58.0	-
log(DOR)		0	-	-	0

2.1 Gender based on a user’s first name

Twitter accounts are uniquely identified by a user id, a number assigned during the creation of the account. Apart from that, a screen name has to be provided (the name after the @-sign) and a username. The latter is often the name of the person in real life. If this is the case, the first name can be used as in indicator for somebody’s gender. Information on the number of men and women with a particular first name in the Netherlands is available in the online Dutch first name database of the Meertens institute; a Dutch Academy of Science institute for language and culture (www.meertens.knaw.nl). An R script was created that cuts out the first part of the username of a Twitter account, i.e. the part before the first space, and checks the occurrence of the number of men and women, with that particular name, in the Meertens institute database. The result is expressed as a number between 0 and 1, indicating the fraction of men of the total number of persons with that name. A name unique to women thus has a score of 0 and a name unique to men has a score of 1. In Table 4 a selection of men’s names and women’s names and their respective score is shown. For names not included in the database the script returned a value of −1 (unknown).

Table 4. Examples of first names and their score provided by the first name database of the Meertens institute.

First name	Score	First name	Score	First name	Score
Aaldert	1	Aafke	0	Anne	0.239
Erik	1	Els	0.003	Emanuele	0.722
Jeroen	0.999	Joana	0	Joan	0.347
Kees	0.996	Katja	0	Kim	0.023
Martijn	0.999	Merel	0	Marti	0.467
Piet	0.999	Petra	0	Pleun	0.302
Vladimir	1	Vivianne	0	Wendel	0.505

When the script was applied to the usernames provided by the 691 selected person accounts, 633 (92%) of the names were found to be registered in the first name database. Since numeric values were produced, including values between 0 and 1, the results were also converted to 0 and 1 below and above specific cut-off values, respectively. This resulted in the findings shown in Table 5.

Table 5. Performance measures for gender classification from first names at various cut-off values. Shaded cells contain highest scores.

Measure	Perspective	cut-off						
		♀ 0	<0.1	<0.2	<0.3	<0.4	<0.5	<0.5
		♂ 1	>0.9	>0.8	>0.7	>0.6	>0.5	≥0.5
Accuracy (%)		38.5	86.5	87.6	88.3	88.7	89.1	89.9
Sensitivity (%)	Male	33.3	88.3	88.5	89.7	90.5	91.0	92.2
	Female	46.1	84.0	86.2	86.2	86.2	86.5	86.5
Precision (%)	Male	99.3	99.4	99.5	98.7	98.4	98.4	98.2
	Female	99.2	98.8	98.8	98.0	98.0	98.0	98.0
F1 (%)	Male	49.8	93.5	93.7	94.0	94.3	94.5	95.1
	Female	63.0	90.8	92.0	91.7	91.7	91.9	91.9
log(DOR)		−0.85	3.68	3.87	4.00	4.08	4.17	4.33

The results in Table 5 demonstrate that very quickly, at low cut-off values, males and females can be classified with great accuracy and sensitivity. An accuracy of nearly 90% is achieved at best. Female sensitivity is always lower than the findings for males indicating that first name classification is more difficult for female names. The precision for both genders starts very high (above 99%) and only slightly drops to 98% with increasing cut-off values. Both F1 and the DOR continue to increase with increasing cut-off values. Both values are highest when the maximum cut-off value is reached: 0.5. This demonstrates that the whole range of values can and should be used. The last column in Table 5 reveals that it pays off to denote persons with an exact value of 0.5 as men. This is not unexpected since half the persons in the sample are males (see Table 1). Table 5 indicates that females are the most difficult to classify gender when using first names.

2.2 Gender based on short biography

Twitter users can also provide a short biography on their user's account webpage. In the biography section they can provide information on their personal life, which may contain clues on their gender. In the persons dataset 465 (67%) provided a short biography, which was either written in Dutch or in English. This text was checked for the occurrence of words referring to the gender-related position in the family, such as (grand)mother of, (grand)father of, son of, daughter of, aunt of, uncle of, etc.; in both Dutch and English. It was found that a total of 154 (33%) bio's contained such words. The absence of gender-related text resulted in many unclassified persons. Hence only an accuracy of 22% was obtained (Table 6). However, the gender information provided by these words was found to be correct to a very high extent; the precision was 96% for males and 100% for females. The effect of many unclassified persons is also reflected in both the F1 and DOR measures. The F1 scores are quite low. The log(DOR) is negative, which is caused by many persons not being classified, resulting in high MU and FU values (see Table 2). This makes clear that using the short biography as a single input for gender classification is not a good idea. In combination with other classifiers and in particular for the identification of females it could, however, be worthwhile because of its high precision.

Table 6. Performance measures for gender classification from short biographies.

Measure	Perspective	Performance
Accuracy (%)		21.7
Sensitivity (%)	Male	25.2
	Female	16.7
Precision (%)	Male	96.3
	Female	100
F1 (%)	Male	39.9
	Female	28.6
log(DOR)		−2.70

2.3 Gender based on Tweet writing style

In cooperation with Dong Nguyen of the University of Twente, the gender classifier developed by her and her coworkers (Nguyen et al. 2013) was used to classify the Twitter users in the dataset. Since this classifier uses tweets, the findings for the users who had written tweets were determined first. A total of 473 of the 844 users (56%) had ever written tweets. For these users the gender (and an indication of their age) was determined. As an additional check, the classifier was subsequently also applied to the remaining (tweet-less) users. For all clarity, this means in our case users for which we could not download any tweets. To our surprise nearly half (48%) of these users got a gender assigned by the Tweetgenie API. Since the API uses the screenname as input and subsequently accesses the user account online, it is unknown to the authors what kind of information is used for this classification. Because of this unexpected finding the results for all users and those who wrote and did not write tweets are included in Table 7. These results demonstrated that users are indeed classified more accurately when tweets are available; compare 73% for users with tweets with 38% for users without tweets. The sensitivity for both genders increases when tweets are available and this increase is higher for females. The findings also demonstrate that the precision of this approach is fairly good but is hardly affected by the presence of tweets. This is surprising and indicates that the classifier is not very good at discerning between truly and falsely classified males or females, respectively. The F1 and DOR reveal fairly high values when tweets are available, especially for females, but not as high as when using first names. Compared with the latter alternatives this classifier is not the best option. Apart from that, we recommend using it only for accounts that have actually written tweets.

Table 7. Performance measures for gender classification from Dutch tweets. Results for all users (691) and users that produced tweets (367) are shown.

Measure	**Perspective**	**All users**	**Users with tweets**	**Users without tweets**
Accuracy (%)		56.6	73.3	37.7
Sensitivity (%)	Male	62.6	75.6	46.7
	Female	47.9	69.7	25.7
Precision (%)	Male	82.8	83.7	81.1
	Female	74.6	73.9	76.6
F1 (%)	Male	71.3	79.4	59.3
	Female	58.3	71.7	38.5
log(DOR)		0.43	1.96	−1.19

2.4 Gender based on profile picture

The fourth option to determine gender is by using the profile picture provided by the user. A total of 661 of the persons in the sample provided a non-default picture. However, before one is able to classify the face(s) on the picture, the picture needs to be standardized to make sure the classifier is trained with the best data available. This means that faces need to be identified, extracted and aligned. These faces are subsequently classified as male or female.

2.4.1 Face extraction and standardization

For face extraction and standardization we used the open source OpenCV software (Bradski 2000) and wrote a Python script. First experiments revealed that it was difficult to identify all faces on the Twitter profile pictures. Next, alignment of the faces, by identifying eyes and/or the nose and mouth proofed challenging as well. To maximize face extracting and alignment we developed the approach shown in the flowchart in Figure 1.

The script was able to identify faces or facial features on 491 (74%) of the 661 useable pictures. Only one of the extracted features did not represent a face; it was the result of the combination of light and shadow caused by a lamp. The number of TP therefore was 490 and the number of FP was 1. From these 490 pictures with actual facial features a total of 516 faces were extracted. Of those faces, 472 were

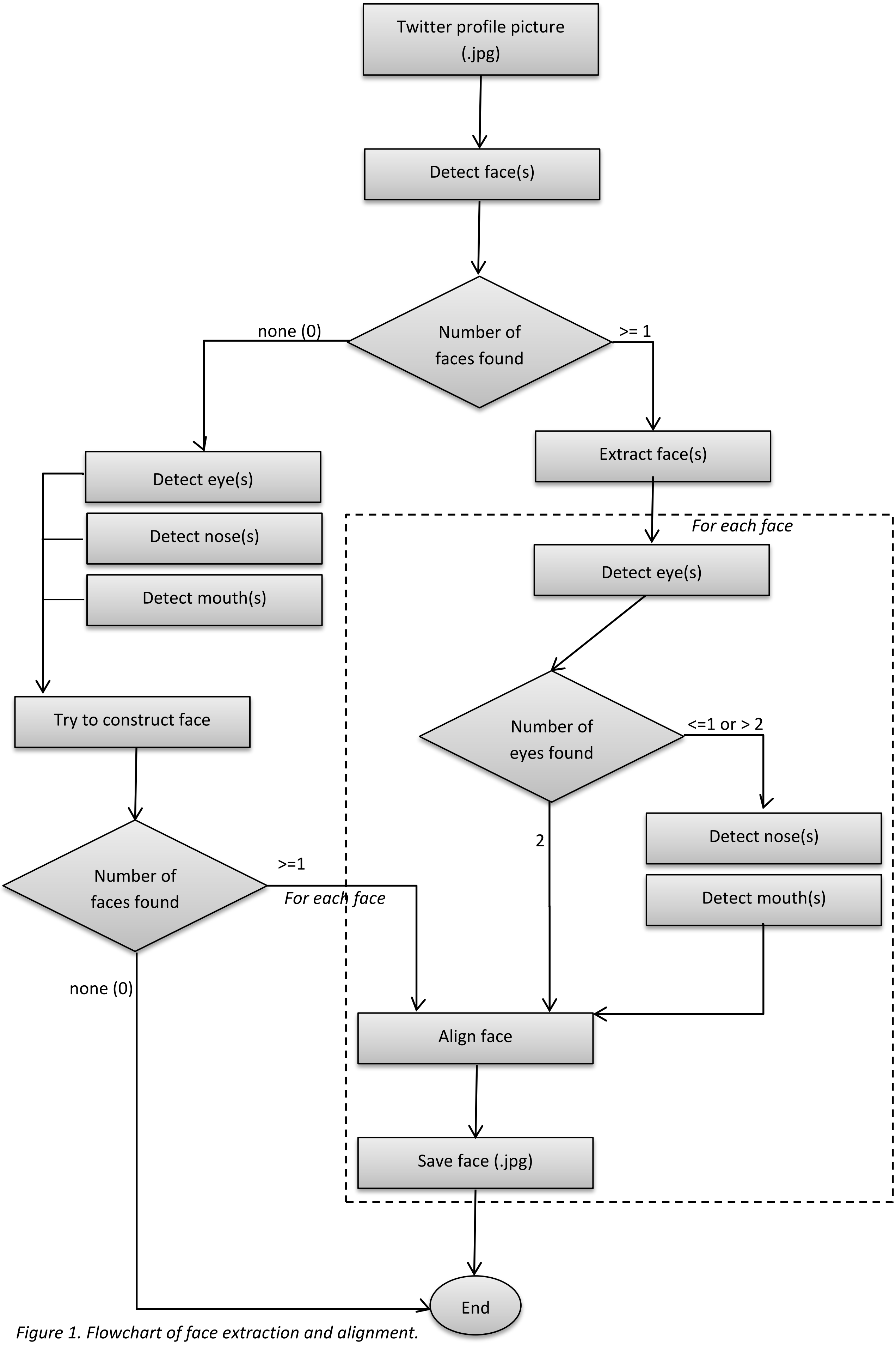

Figure 1. Flowchart of face extraction and alignment.

derived from pictures containing a single face. From the remaining 18 pictures two or more faces were obtained, indicating that more than one person or multiple copies of the same person were included. The maximum number of faces extracted by the script from a single picture was four. Of the 661 pictures, 155 pictures were found to contain a face or facial features that were not recognized by the script. The number of FN was therefore 155. The pictures on which no face was shown, the TN, were found to be 15; for clarity, these pictures did not have the default egg displayed on them (see above). If we consider the identification of 1 or more faces as a single positive outcome, Table 8 gives the facial identification performance of the script.

Table 8. Performance measures for face extraction from profile pictures by the Python script. Results of the 661 non-default pictures are shown and the identification of 1 or more faces is considered as a single positive outcome.

Measure	**Perspective**	**Performance**
Accuracy (%)		76.4
Sensitivity (%)	Face(s)	76.0
	No face	93.8
Precision (%)	Face(s)	99.8
	No face	8.8
F1 (%)	Face(s)	86.3
	No face	16.1
log(DOR)		3.86

This demonstrated that the script was very precise; nearly all objects extracted were faces. There is room for improvement though, as is shown by the sensitivity and accuracy of the script. Not all faces or facial features were identified; around 24% were missed. This is remarkable as the script checked for a whole range of facial features at various levels of detail (see Figure 1). These findings indicate the diversity of the facial features included on Twitter pictures. It will be challenging to improve this.

Next, the 516 identified faces were standardized as much as possible. If the face was directly identified, as shown on the right path of the flowchart in Figure 1, both eyes were attempted to be detected. If that was successful, a straight line was drawn between the centers of those eyes. This line was subsequently used to rotate the face in such a way that the eyes were aligned in a horizontal fashion. Faces, on which none or only one eye was detected, were additionally checked for the occurrences of a nose and a mouth. If both features were found, a straight line was drawn between these features. This line was subsequently used to rotate the face such that the nose and mouth were aligned in a vertical way. Faces, on which none or only one eye and no combination of nose and mouth were detected, were not aligned. Pictures, on which no faces were detected, were always checked for the occurrences of eyes, noses and mouths; this is shown on the left path of the flowchart in Figure 1. Depending on the features revealed, alignment was attempted as described above. The resulting faces (516) were supplemented with 11 pictures on which full head shots of non-extracted faces of high quality were present, and subsequently used for gender classification.

2.4.2 Gender classification from facial images by support vector machines

A total of 527 standardized and annotated facial images were available for gender classification. Each image consisted of three 400×400 matrices, each matrix containing the relative color intensity of either the red (R), green (G) or blue (B) primary, and each cell containing the relative color intensity of the pixel. The three RGB matrices were first condensed to a single grayscale matrix, using weights 0.30 (R), 0.59 (G) and 0.11 (B) (ITU-R 2011). Dimensionality was reduced by singular value decomposition, further reducing the matrix to a vector of 400 singular values—the square root of the eigenvalues of the covariance matrix (see e.g. Hogben 2007).

The set of 527 images was randomly split into a training set (70%) and a test set (30%). The training set was used to train a support vector machine (SVM; Hastie et al. 2009) the relationship between the singular values and the annotated gender. The test set was used to predict the gender from the singular values by the trained SVM. The test set thus contained both an annotated and a predicted gender for each facial image. From these a contingency table (or confusion matrix) could be constructed and the performance measures calculated. The procedure of splitting, training and testing was repeated 50 times to obtain a rough estimate of the variability in the performance.

SVMs can perform nonlinear classification by projecting the input into a higher dimensional space where males and females become linearly separable. To this end the choice of a nonlinear kernel function becomes important. We reran the procedure with a polynomial, radial and sigmoid kernel in addition to a linear kernel to see which gave the best results. To prevent overfitting, additionally the optimal number of singular values was determined by rerunning the procedure with different numbers of singular values (ranging from 1 to 400).

The radial kernel with 327 singular values gave the highest diagnostic odds ratio (Figure 2). Although the DOR was not as high as the previous methods, it was considerably higher than random guessing based on the sex ratio in the training set (Figure 3). All performance measures are given in Table 9. Most striking is the low sensitivity for women.

Table 9. Performance measures for gender classification from facial images. Pooled results from 50 replicates of 158 test cases each.

Measure	Perspective	Performance
Accuracy (%)		58.6
Sensitivity (%)	Male	90.3
	Female	16.0
Precision (%)	Male	59.1
	Female	55.0
F1 (%)	Male	71.4
	Female	24.8
log(DOR)		0.57

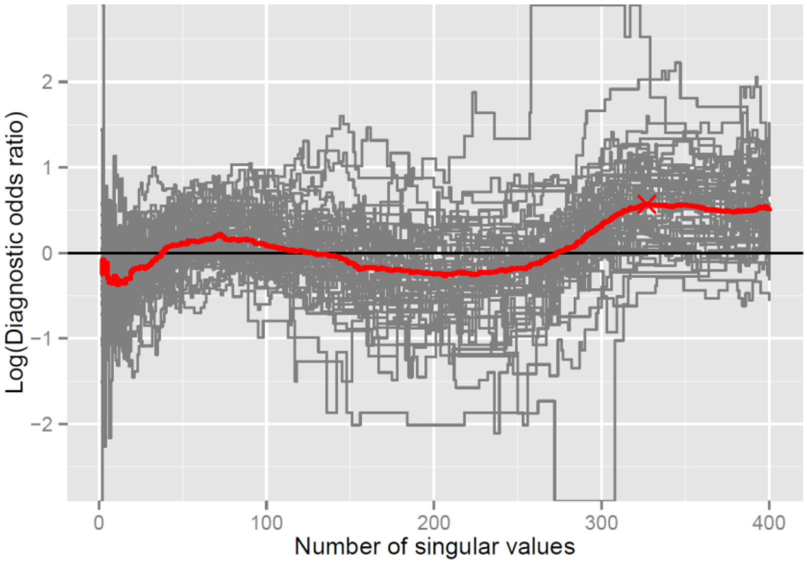

Figure 2. Effect of the number of singular values on the diagnostic odds ratio (DOR) using an SVM with a radial kernel. Red line shows results when 50 replicates (gray lines) are pooled; red cross shows highest DOR; black line shows null hypothesis (DOR = 1).

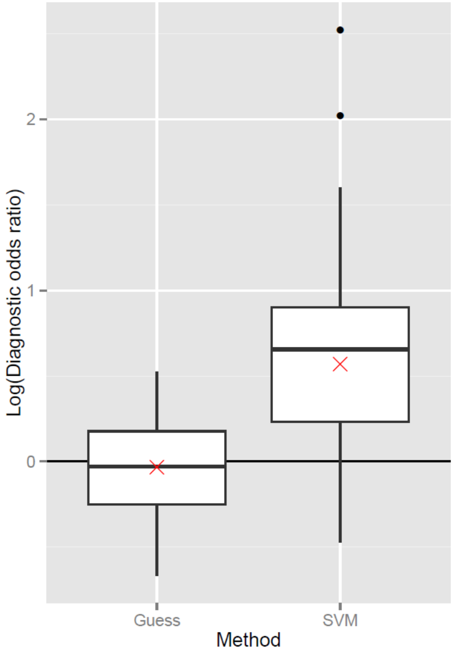

Figure 3. Distribution of diagnostic odds ratio among 50 replicates using either random guessing based on the sex ratio in the training set or an optimized SVM (radial kernel and 327 singular values). Red crosses shows results when replicates are pooled; black line shows null hypothesis.

2.5 Gender based on combining classifiers

From the above it is clear that gender classification based on first names performed best; a maximum log(DOR) of 4.33 was obtained. However, this performance could be improved by combining all four information sources provided by a Twitter account. This resulted in a maximum log(DOR) of 7.02. To achieve this, gender classification based on keywords in the short biography were used as a first step; this information source is very precise to females. After this step, of the 844 users, 689 had no gender assigned. These were subsequently subjected to first name gender classification. Hereafter, 153 users remained that had no gender assigned. Next, tweet writing style classification was applied to these, resulting in 29 users to which no gender had been assigned. To these, the results of the facial image gender classification were added, resulting in 20 users to which no gender was assigned. As a last step, all these users were classified as males. The end results are summarized in Table 10.

Table 10. Performance measures for the gender classification from the combination of short biography, first names, Dutch tweets and profile pictures are shown.

Measure	**Perspective**	**Performance**
Accuracy (%)		96.5
Sensitivity (%)	Male	98.8
	Female	93.3
Precision (%)	Male	95.5
	Female	98.1
F1 (%)	Male	97.1
	Female	95.6
Log(DOR)		7.02

3. Auxiliary information from a matched source

In this section we explore the possibilities to extract additional auxiliary information by matching Twitter with another social media source, i.e. LinkedIn. We used 836 of the 844 still existing Twitter accounts (Section 2) as a frame for testing our method of profiling. These 836 accounts all had a screen name, username, and a unique user id. Of these accounts 576 (69%) provided a short description and 794 (95%) had location information.

We used this information to obtain additional auxiliary information from an associated LinkedIn account, such as gender, age, level of education, type of industry, type of job, location of resident, etc.

Statistics Netherlands already applied web scraping techniques to retrieve data from Internet sources for statistics (ten Bosch and Windmeijer 2014). Based on this experience, we developed a robot which was able to match Twitter usernames with associated LinkedIn accounts, and was able to search and retrieve auxiliary information from these LinkedIn accounts.

The results were promising. For 568 Twitter accounts (68%) one or more matches were found on LinkedIn (Figure 4). For the other 268 Twitter accounts (32%) no LinkedIn-accounts were found. Presumably, people do not use their real name on Twitter or do not have a LinkedIn account at all. Out of the 568 matches, 215 matches referred to one LinkedIn-profile only (single hits). The other 353 matches referred to multiple LinkedIn profiles per Twitter account, which led to a new challenge: to find the best match. In addition, single hits might not be correct and have to be examined.

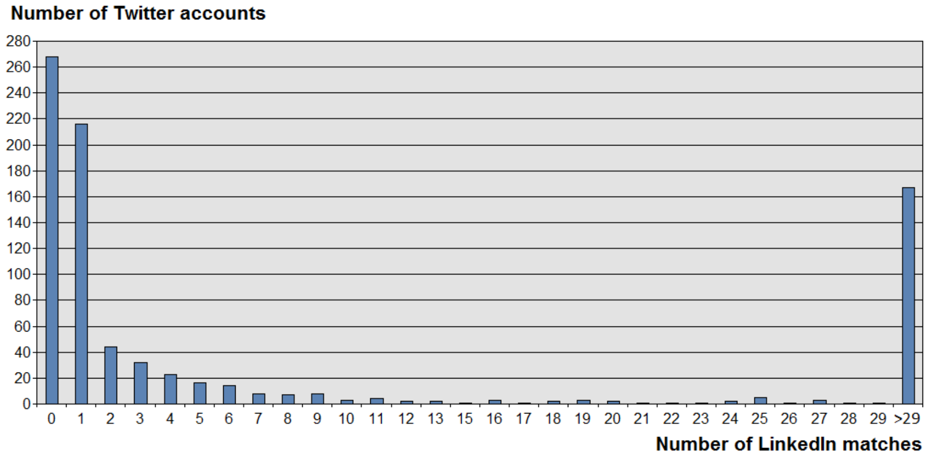

Figure 4. Frequency distribution of the number of LinkedIn matches per Twitter account.

For both types of hits, single hits as well as multiple hits, a score function was introduced. This function calculates a score from 0 to 100 for each match that is found. The matching score S_{ij} between Twitter account i and matched LinkedIn account j is defined as:

$$S_{ij} = c_i \sum_k w_k s_{ijk},$$

where c_i is a correction factor, w_k is a weighting factor, and s_{ijk} is the proportion of all m_i scores on feature k of Twitter account i that are accounted for by match ij. The proportion is expressed as percentage. Up to 25 scores are calculated per Twitter account, because this is the maximum number of matches that is displayed to the robot on the first page of the LinkedIn search results. Thus:

$$s_{ijk} = \frac{b_{ijk}}{\sum_{j=1}^{\min(m_i,25)} b_{ijk}} \times 100\%,$$

where b_{ijk} is the score of match ij on feature k. Each match is scored on three features: name ($k = 1$), location ($k = 2$) and Twitter description ($k = 3$). For the latter the longest common substring algorithm is used, it checked all the words from the Twitter description with the descriptive fields of LinkedIn such as type of industry, location of resident, job description, name of employer, description of education and summary, and kept track of the maximum. In Table 11 some examples of b_{ijk}, the score of match ij on feature k, are shown. For features $k = 1$ and $k = 2$, these were chosen intuitively, for $k = 3$ the score is the number of matching words (ignoring filler words).

Table 11. Examples of b_{ijk}, the score of match ij on feature k.

Feature k	**Match**	b_{ijk}
1 (name)	Null	0
	Full name	100
	Surname	5
	First name	5
2 (location)	Null	0
	Delft	100
	Rotterdam, Nederland	100
	Nederland	30
	Netherlands	30
	Sweden	100
	Hamburg	100
	Sweden	100
3 (description)	Null	0
	Account manager	2
	Marketing, Leerdam EOC	3
	Medical Emergency	2
	Consult, Coolblue	2
	zzp, java, perl, Hoorn	4
	Radio	1

The three relative scores are weighted by $w_k = \{0.55, 0.10, 0.35\}$ $(\sum_k w_k = 1)$. Weights were chosen by trial and error. The score S_{ij} is downgraded by a correction factor c_i when more than 25 LinkedIn accounts are matched:

$$c_i = \begin{cases} 1 & \text{if } m_i \leq 25 \\ \frac{25}{m_i} & \text{if } 25 < m_i \leq 100. \\ 0 & \text{if } m_i > 100 \end{cases}$$

For example, if 50 LinkedIn accounts are found for a Twitter account while only 25 scores, shown on the first page, are calculated, the correction factor c_i, for these 25 scores, is 0.5. Figure 5 shows a frequency distribution of the highest score of the 568 Twitter accounts that had at least one LinkedIn match. For single hits, there is only one candidate LinkedIn account. For multiple hits, the LinkedIn profile with the highest score ($\max_j S_{ij}$) was selected—if not more than 100 matches were found—as the one that we presume belongs to the Twitter account.

To get a feeling for the quality of the matching method, we manually compared the content of the Twitter account to that of the LinkedIn account. Because of limited resources, only the profiles with a score of 50 or higher were analyzed (the yellow right area in Figure 5)

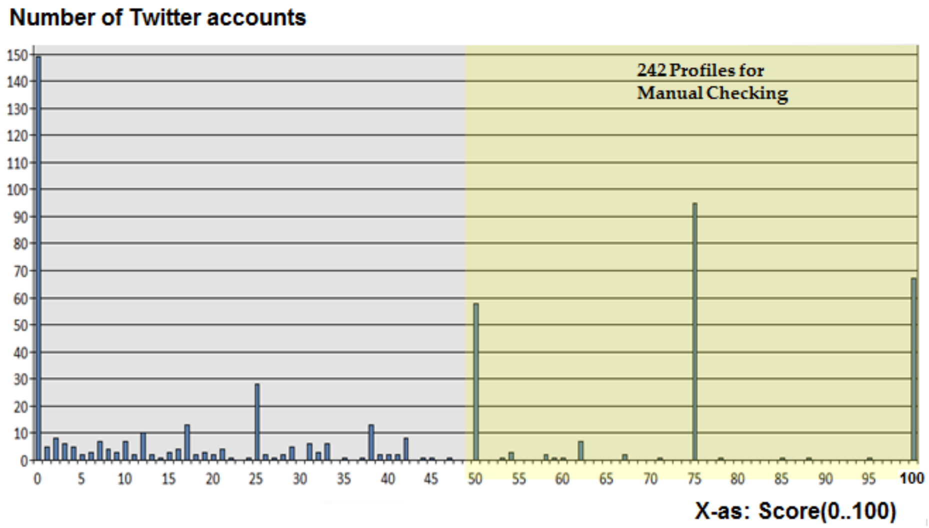

Figure 5. Frequency distribution of the highest score per Twitter account.

The manual checking process, of 242 profiles, consisted of two parts: comparing profile pictures and comparing tweet and LinkedIn content. The latter contained variables such as type of industry, location of resident, job description, name of employer, level of education, description of education and parts of the summary. On the basis of this, we were able to estimate which portion of the matches were true, false or undetermined. We concluded that the score gives a reasonable indication for the probability that the match is correct (Table 12).

Other characteristics such as gender and age were not directly obtainable from LinkedIn. However, gender could potentially be derived from the combination of job,

Table 12. Manual checking: probability the best match is a true positive (a correct match), a false positive and undetermined for profile scores ≥ 50.

Score	**True positive (%)**	**False positive (%)**	**Undetermined (%)**
> 75	95.8	2.8	1.4
60–75	67.9	18.9	13.2
50–59	58.5	29.2	12.3

interest and the profile summary. Age a could be derived from a combination of the year in which the education was started t_0, the year of graduation t_1, the current year t_2, and the educational level e:

$$a = \begin{cases} a_{0e} + t_2 - t_0 & \text{if } t_0 \text{ is known} \\ a_{1e} + t_2 - t_1 & \text{if } t_0 \text{ is unknown but } t_1 \text{ is'} \end{cases}$$

where a_{0e} and a_{1e} is the age at which educational level e is usually started and finished, respectively (Table 13). A combination was found in 121 cases of the exact matches with a score higher than 50. In 49 of these cases either start year or year of graduation from secondary school is available. From all the information that we retrieved we believe that this gives the best estimation of age. In the other 72 cases we estimated the age from the year a person started a higher education. For this we used the Dutch educational type and the starting year or the year of graduation (whatever was available).

Table 13. Age at which education e is usually started a_{0e} and finished a_{1e}.

$\boldsymbol{e}$	$\boldsymbol{a_{0e}}$	$\boldsymbol{a_{1e}}$
Lower general secondary education	12	16
Higher general secondary education	12	17
Pre-university education	12	18
Intermediate vocational education	16	18
Higher vocational education	17	20
Bachelor	18	21
Master	21	23
PhD	23	28

Figure 6 shows the age distribution of this sample in relation to the age distribution of the Dutch population in 2015. Although it has to be noted that the size of the resulting sample is rather small, so care should be taken to draw conclusions here, it looks like persons in their late twenties and early thirties are strongly overrepresented in the matched Twitter-LinkedIn sample, whereas children and elderly are not represented at all.

The results of the age detection support the idea that the matched Twitter-LinkedIn sample is not representative for the complete population. This is obvious since the method of using public LinkedIn accounts as described in this section has some natural biases. First, we used publically shared information only, which of course creates a bias in favor of people publicly exposing their profile. Second, even for the set of public available LinkedIn profiles, the information richness per person varies.

From the experiments in this section we conclude that we can use multiple social media sources such as LinkedIn to obtain auxiliary variables. An example is age for which we used profile information such as level and type of education, the starting year and year of graduation.

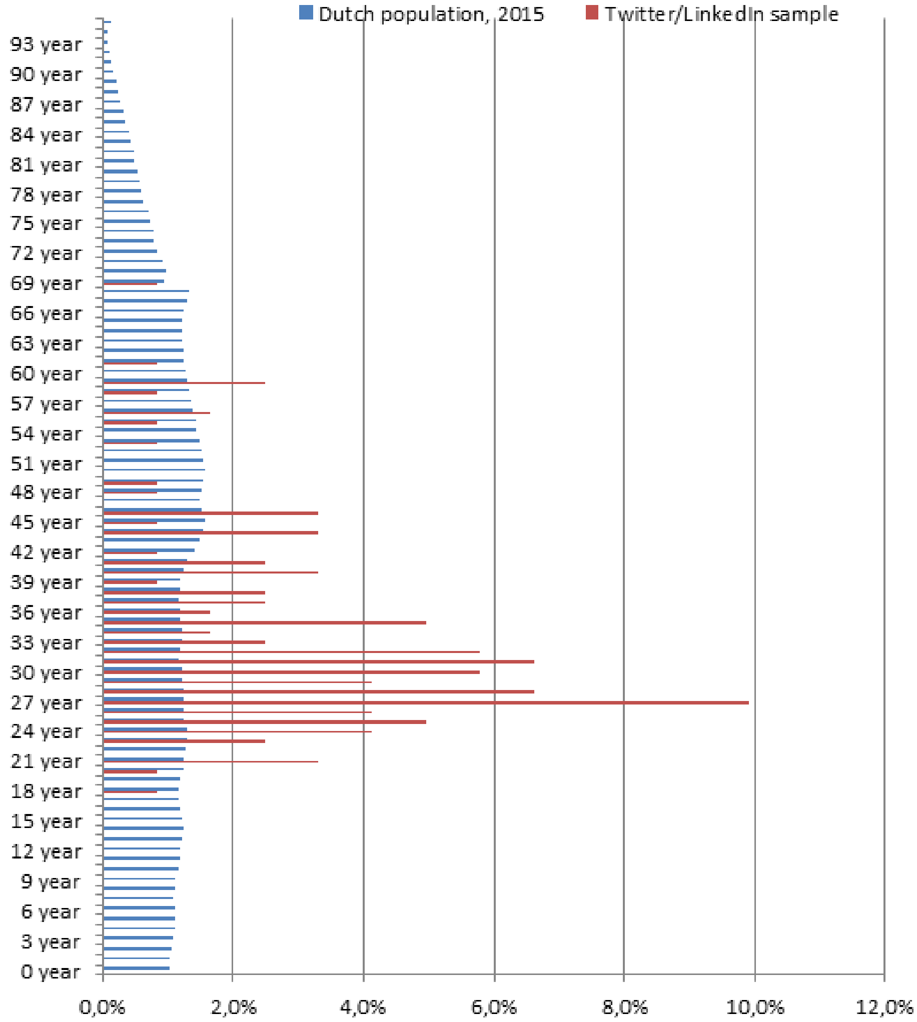

Figure 6. Age distribution of Twitter-LinkedIn sample and Dutch population.

4. Discussion

Big data are acknowledged as potential data sources that should be taken advantage of in the production of official statistics (Daas et al. 2015). Here we focused on a methodological challenge: their representativeness relative to a target population. Unlike in sample surveys, the mechanism generating big data is not a probability sample. As a result, big data may cover a selective part of the target population. Auxiliary information explaining the missingness could be used to quantify and correct for this selectivity, but linking this information from registers is often not feasible. Here we have explored the possibilities of i) extracting auxiliary variables from a big data source itself and ii) obtaining them from another source. Using Twitter as a case for the first study, we have shown that the commonly used auxiliary variable gender can be determined using the user name, the short biography, public tweets and the profile picture. For the second study, associated LinkedIn accounts were used to obtain several additional characteristics. It was also found that only a part of the Twitter accounts could be linked.

In Twitter, gender classification performed best on first names. This performance could be improved by optimally combining all four information sources provided by a Twitter account. Here, information provided by the short biography was applied first, followed by first name, tweet writing style and facial image gender classification. The small portion of users remaining to which no gender was assigned was subsequently classified as men. Gender classification performed worst on facial images, although the support vector machine still performed better than random guessing. Computer vision in general and gender classification from facial images in particular is an active area of research (see e.g. Ng et al. 2012). Most studies, however, use databases with standardized faces. Twitter images are much more heterogeneous in composition, lighting conditions etc. Extracting additional features, expanding the training set and applying alternative classifiers (e.g. random forests) provide ample opportunity for improvement. The approach followed for age determination from the tweet writing style (see Nguyen et al., 2013) produces a single estimate. Alternatively, age could also be derived from the first name of the user (see Figure 7 for examples). In this case, the results are probability distributions for the age of persons with these names.

More auxiliary information can be retrieved when Twitter accounts are matched with other sources, such as LinkedIn accounts. We found that about two thirds of the Twitter accounts could be matched to one or more LinkedIn accounts. A matching score function was developed to indicate the probability that a match is correct. About 8% of the Twitter accounts could be matched to a LinkedIn account that was confidently of the same person. Here the volume of big data may come in useful: matching a million Twitter accounts would still leave an appreciable eighty thousand accounts with auxiliary information. These accounts are, however, more likely to represent a selective part of the Twitter population: the part of the users' active both on Twitter and LinkedIn. Matching of additional sources will be necessary to improve coverage.

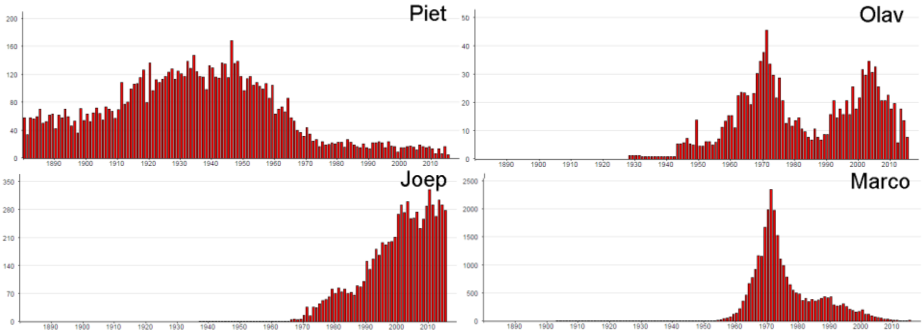

Figure 7. Popularity over time of the first names of four of the authors in the database of the Meertens institute (As a Dutch first name, Quan is too rare to occur in the public database). The distributions indicate differences in the popularity of the names for men born in the Netherlands between 1880 and 2015. As such, a first name indicates a particular age distribution for persons with such a name.

Comparing the distribution of auxiliary variables between the profiled sample and the population register reveals the following. In the Twitter sample males are strongly overrepresented. This indicates that men are more active on Twitter. In the Twitter-LinkedIn matched sample, persons in their late twenties and early thirties are strongly overrepresented, whereas children and elderly are not represented at all. The latter suggest an even stronger selection when the combination of two sources is used.

Provided that background characteristics, such as gender and age, can be measured without error and correlate with the variable of interest, such as the sentiment in messages, this information could be used to improve the accuracy of social media based findings. Possible applications for this are a sentiment based indicator (Daas et al. 2014) and a ‘feeling of safety’ indicator (Steffens 2016) which both use social media.

Acknowledgements
We thank Dong Nguyen for allowing us to use the API of tweetgenie.nl and Bart Buelens for stimulating discussions and useful comments on an earlier version of this paper.

5. References

Bosch, O. ten and D. Windmeijer, 2014. On the use of internet robots for official statistics. UNECE meeting on the Management of Statistical Information Systems (MSIS), Dublin, Ireland.

Bradski, G., 2000. The OpenCV library. Dr. Dobb's Journal of Software Tools.

Buelens, B., P. Daas, J. Burger, M. Puts and J. van den Brakel, 2014. Selectivity of big data. Discussion paper 201411. Statistics Netherlands, The Hague / Heerlen, the Netherlands.

Buelens, B., J. Burger and J. van den Brakel, 2015. Predictive inference for non-probability samples: a simulation study. Discussion paper 201513. Statistics Netherlands, The Hague / Heerlen, the Netherlands.

Daas, P.J.H. and M.J.H. Puts, 2014. Social media sentiment and consumer confidence. European Central Bank Statistics Paper Series No. 5, Frankfurt, Germany.

Daas, P.J.H., M. Roos, M. van de Ven and J. Neroni, 2012. Twitter as a potential data source for statistics. Discussion paper 201221, Statistics Netherlands, The Hague/Heerlen, the Netherlands.

Daas, P.J.H., M.J.H. Puts, B. Buelens and P.A.M. van den Hurk, 2015. Big data as a source for official statistics. *Journal of Official Statistics* 31: 249–269.

Eurostat, 2013. Internet access and use in 2012. Eurostat newsrelease. Located at: http://epp.eurostat.ec.europa.eu/cache/ITY_PUBLIC/4-18122012-AP/EN/4-18122012-AP-EN.PDF

Glasson, M., J. Trepanier, V. Patruno, P. Daas, M. Skaliotis and A. Khan, 2013. What does "big data" mean for official statistics? Paper for the High-Level Group for the Modernization of Statistical Production and Services.

Hastie, T., R. Tibshirani and J. Friedman, 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Second edition. Springer, New York, NY, USA.

Hildebrandt, M. and S. Gutwirth, 2013. *Profiling the European Citizen. Cross Disciplinary Perspectives*. Springer, Dordrecht, the Netherlands.

Hogben, L., 2007. *Handbook of Linear Algebra*. Chapman and Hall/CRC, Boca Raton, FL, USA.

ITU-R, 2011. Recommendation BT.601-7. International Telecommunication Union, Geneva, Switzerland.

Ng, C.B., Y.H. Tay and B.M. Goi, 2012. Vision-based human gender recognition: a survey. arXiv:1204.1611.

Nguyen, D., R. Gravel, D. Trieschnigg and T. Meder, 2013. "How old do you think I am?": A study of language and age in Twitter. In: Proceedings of the seventh international AAAI conference on weblogs and social media. AAAI Press, Palo Alto, CA, USA.

Silver, N., 2012. *The Signal and the Noise: Why So Many Predictions Fail—but Some Don't*. Penguin, New York, NY, USA.

Shaikh, S.A., 2011. Measures derived from a 2 x 2 table for an accuracy of a diagnostic test. *Journal of Biometrics and Biostatistics* 2(5).

Steffens, P. (2016) Measuring safety using social media: an applied sentiment analysis through the use of text mining. MSc thesis. University of Maastricht, Maastricht, the Netherlands.

Explanation of symbols

Empty cell	Figure not applicable
.	Figure is unknown, insufficiently reliable or confidential
*	Provisional figure
**	Revised provisional figure
2015–2016	2015 to 2016 inclusive
2015/2016	Average for 2015 to 2016 inclusive
2015/’16	Crop year, financial year, school year, etc., beginning in 2015 and ending in 2016
2013/’14–2015/’16	Crop year, financial year, etc., 2013/’14 to 2015/’16 inclusive

Due to rounding, some totals may not correspond to the sum of the separate figures.

Publisher
Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague
www.cbs.nl

Prepress
Statistics Netherlands, Studio BCO

Design
Edenspiekermann

Information
Telephone +31 88 570 70 70
Via contact form: www.cbsl.nl/information