Creating Extended Gender Labelled Datasets of Twitter Users

Marco Vicente[1,2], Fernando Batista[1,2], and Joao P. Carvalho[1,3]

[1]L[2]F – Spoken Language Systems Laboratory, INESC-ID Lisboa
[2]Instituto Universitário de Lisboa (ISCTE-IUL), Lisboa, Portugal
[3]Instituto Superior Técnico, Universidade de Lisboa, Portugal
m.vicente.pt@gmail.com, {first_name.last_name}@inesc-id.pt

Abstract. The gender information of a Twitter user is not known *a priori* when analysing Twitter data, because user registration does not include gender information. This paper proposes an approach for creating extended gender labelled datasets of Twitter users. The process involves creating a smaller database of active Twitter users and to manually label the gender. The process follows by extracting features from unstructured information found on each user profile and by creating a gender classification model. The model is then applied to a larger dataset, thus providing automatic labels and corresponding confidence scores, which can be used to estimate the most accurately labeled users. The resulting databases can be further enriched with additional information extracted, for example, from the profile picture and from the user location. The proposed approach was successfully applied to English and Portuguese users, leading to two large datasets containing more than 57K labeled users each.

Keywords: Gender classification, Twitter users, Gender database, Text Mining

1 Introduction

Existing social networking services provide means for people to communicate and express their feelings in a easy way. Such user generated content contains clues of user's behaviors and preferences, as well as other metadata information that is now available for scientific research. Twitter, in particular, has become a relevant source for social networking studies, mainly because: it provides a simple way for users to express their feelings, ideas, and opinions; makes the user generated content and associated metadata available to the community; and furthermore provides easy-to-use web interfaces and application programming interfaces (API) to access data. For many studies, the different attributes about a user may be relevant. However, Twitter registration does not explicitly include relevant information such as, for example, gender (not even optionally). For that reason, many previous studies involving Twitter had to rely on small manually labelled datasets of users. Manual labelling represents a labor-intensive task and is a very demanding challenge when analysing social media given the usual huge number of users.

Table 1. Twitter labelled datasets reported in the literature.

Study	Users	Tweets	Languages	Geography
Rao et al. (2010) [18]	1000	405k	English	India
Burger et al. (2011) [5]	183729	4.1M	Several	
Bergsma et al. (2013) [3]				
Liu et al. (2012) [11]	400	N/A	English	Canada
Bamman et al. (2012) [2]	14464	9.2M	English	United States
Deitrick et al. (2012) [7]	N/A	3031	English	
Fink et al. (2012) [8]	11155	18.5M	English	Nigerian
Miller et al. (2012) [14]	3000	N/A	English	
Al Zamal et al. (2012) [1]	400	N/A	English	Canada
Liu et al. (2013) [12]	8000	8M	English	
Ciot et al. (2013) [6]	8118	N/A	Several	
Kokkos et al. (2014) [10]	N/A	10000	English	
Ugheoke (2014) [19]	1000	N/A	English	
Helteren et al. (2014) [9]	600	N/A	Dutch	
Nguyen et al. (2014) [15]	3000	N/A	Dutch	
Van Zegbroeck (2014) [20]	8791	N/A	Flemish	
Vicente et al. (2015) [23]	1464		English, Portuguese	

The creation of Twitter datasets is commonly reported in the literature, and researchers have built databases of Twitter users for many geographic regions and languages, including English [13, 17] and Portuguese [4]. Due to the above reason, most of the reported databases are not labelled with user attributes like gender or age (user age was also not available on Twitter until late 2015). Previous studies reported the task of labelling users with their gender to be demanding, labor-intensive and in many cases not reusable. Table 1 presents a summarised list of labelled datasets reported in the literature, revealing that most of the studies use small labelled datasets when compared to the number of existing users. These studies involve several languages but in the reported literature, English represents 66.7% of the users, Portuguese represents 14.4% and Spanish represents about 6%. In [18], 1000 profiles were manually annotated through the gender/name association using the Twitter profile information (*user name* and *screen name*). Burger et. al [5] report the most extensive database, which was created by following the blogging website links available in the profile of Twitter users, and extracting the gender from the corresponding profiles. To evaluate the accuracy of their method, the authors randomly selected 1000 Twitter users and manually validated them. Only 15% of the sample had explicit gender information. In this case, filtering only Twitter users with blogs may bias the dataset, but also filters bots and spammers. Liu et al. [12] labelled their data using the Amazon Mechanical Turk platform, a platform developed for the distribution of tasks to human workers where each human intelligence task (HIT) is performed by an individual in exchange for a small fee. The reliability of such method is uncertain, even when the same task is performed by more than one person. In [5], the accuracy of Amazon Mechanical Turk human gender

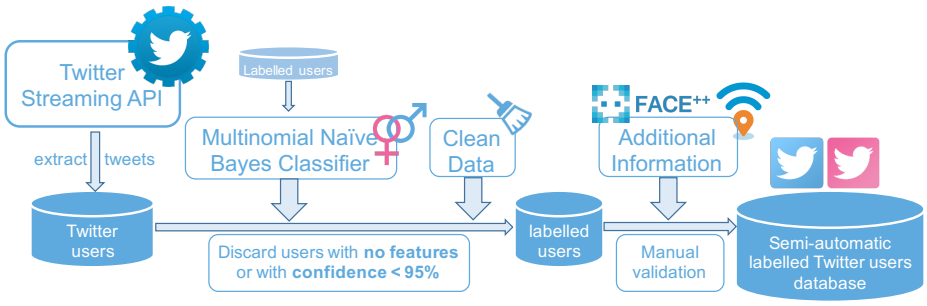

Fig. 1. Semi-automatic gender labelled dataset creation diagram.

classification was only of 68.7%, when averaged across workers. In [1, 6, 7, 10, 11, 14–16, 19, 21], Liu et. al manually labelled users to produce datasets, observing either *user name*, *screen name*, profile picture, tweets or a combination of those attributes. Information available in social media profiles such as Facebook and LinkedIn and associated blogging websites, when provided by Twitter users was also used by [15] and [21, 22].

This paper describes a method for creating extended gender labelled datasets in a semi-automatic fashion. The proposed methodology is language independent, but the focus was currently given to Portuguese and English users. Based on the proposed methodology, two extended labelled datasets of English and Portuguese Twitter users have been created, which can be used, not only to provide additional information for further processing stages, but also to create gender models based on the users' generated content and profile information. This paper is structured as follows: Section summarises the whole process of creating gender labelled datasets, while sections 2 and 3 provide details about two major stages: the creation of core extended labelled datasets, and the enrichment the datasets previously created, respectively. Section 5 describes the data validation. Finally, Section 6 summarises our work and presents the conclusions.

2 Proposed Approach

The approach to create extended labelled datasets is depicted in Figure 1. The first step in the pipeline consists of extracting data from the Streaming API. It involves restricting tweets to a given geolocation, to the set of languages under consideration, and also involves filtering the data in order to remove undesirable users, such as companies and chat bots.

The second step uses a small dataset of labelled users in order to create a gender classification model. The model is based on a set of features extracted from unstructured information found on two profile attributes: *user name* (up to 15 characters), *screen name* (up to 20 characters). The feature extraction process, detailed in [21, 23], considers a number of normalisation steps, such as: *repeated vowels* (e.g.: "eriiiiiiiiic"→"eric"), and *leet speak* (e.g.: "3ric"→"eric").

After finding one or more names, the following elements are addressed in each feature: "case", "boundaries", "separation" and "position". E.g.: Considering the *screen name* "johnGaines", three names can be extracted: "john", "aine" and "ines". The name "aine" has no valid boundaries, since is preceded and succeeded by alphabetic characters. The feature found is weak and the size of the name is lower than the previously defined threshold. Consequently, the name is discarded. The name "ines" has a valid end boundary, as it is not succeeded by alphabetic characters. Finally, the name "john" has a valid end boundary and starts at the beginning of the screen name. Different thresholds have been defined for specific features, e.g. names with such type of boundary (valid end boundary) and such position (start of the screen name) must contain at least 3 characters. At the end of this process 192 features are extracted, including examples such as: "male_name_correct_beginning_separation_and_case", "female_name_beginning_no_separation" [21]. The classification model is then applied to the large dataset, where users having a classification accuracy below a given threshold may be discarded in order to minimise the number of classification errors and improve data quality. The procedure may include optional manual steps to remove other undesirable users that were not detected previously.

Finally, the dataset created can be enriched with additional information that includes, for example, attributes derived from the profile picture and from the user location.

In the scope of this paper, two datasets of Twitter users have been created: a dataset of English users, and a dataset of Portuguese users. The English dataset was extracted from one year of tweets, collected from January 2014 to December 2014 using the Twitter *streaming/sample* API. The data was restricted to active users that have produced at least 100 tweets in English language, either in the United Kingdom or in the United States. From those we kept around 100K users. The Portuguese dataset is the full dataset of the data described in [4], and corresponds to a database of Portuguese users, restricted by users that have tweeted at least 100 tweets in Portuguese language, geolocated in the Portuguese mainland. After creating the datasets we have partially validated the data in order to assess its quality, and split it into *train*, *development* and *evaluation* subsets. The Twitter *streaming/sample* API is technically limited to about 1% of the actual public tweets, but since the amount of geolocated tweets filtered for the Portuguese language within the country geographical area is under 1% of the total number of tweets, the limitations imposed by Twitter are not relevant.

3 Dataset with Core Gender Labels

As previously stated, we have automatically produced two core datasets containing gender labels, based on 192 features extracted from the *screen name* and from the *user name*. This approach was applied to both English and Portuguese users leading to two datasets. Table 2 and Figure 2 present the distribution of the users according to the number of features they positively trigger, revealing a higher portion of English users that did not trigger any of the features (42%) and

Table 2. Number of users that have triggered a given number of gender features.

Dataset	number of users	no features		1 to 10 features		> 10 features	
English	100000	27110	27%	65559	66%	7331	7%
Portuguese	105450	44559	42%	57440	55%	3451	3%

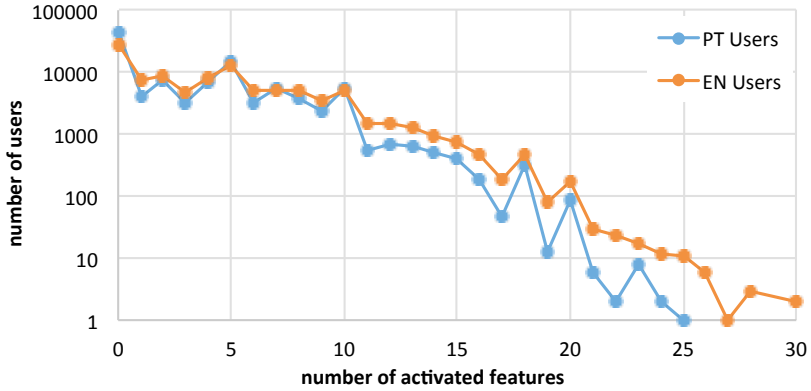

Fig. 2. Automatic Gender Classification - Features per users.

that were, therefore, discarded from the data. The considerably high proportion of discarded users signal two situations: i) the profile of such users information provide little or no clues for gender detection; ii) the feature set can be improved in order to take into account other possible gender detection clues.

In order find the profile attribute that mostly contributes with features to the gender classification task, we have analysed the distribution of features that were extracted either from the *screen name* or from the *user name*. Table 3 shows the obtained results, where columns *user name* and *screen name* represent users activating only features extracted from the corresponding attribute. Results reveal that the two attributes are equally relevant for gender detection.

Based on the extracted features, we have applied supervised machine learning in order to automatically guess the gender label of each user. Different methods have been tested, including: Naive Bayes variants, Logistic Regression, Support Vector Machines, Fuzzy c-Means clustering and k-means, but Multinomial Naive Bayes (MNB) turned out to achieve the best performance [21]. We have used two existing MNB models for English and Portuguese gender classification, previously created from the smaller datasets manually labelled with gender [23]. In the absence of any previously annotated datasets, an alternative approach would be to use an unsupervised gender classification procedure based on Fuzzy C-means clustering [21], which performed almost as well as MNB (96% classification accuracy). After the automatic gender classification stage, users with no features or with features, but classified with a confidence score lower than 95% were discarded in order to minimise the number of classification errors. All the

Table 3. Number of users that have triggered gender features per profile attribute.

Dataset	none		*user name*		*screen name*		both	
English	27110	27%	20845	21%	20580	21%	31465	31%
Portuguese	44599	42%	17776	18%	18443	17%	24672	23%

Table 4. Some of the gender indicative words.

English		Portuguese	
Male	Female	Male	Female
father	mother	pai	mãe
boy	girl	rapaz	rapariga
boyfriend	girlfriend	namorado	namorada
grandfather	grandmother	avô	avó

remaining users were added to a dataset, as well as the text of their 100 most recent tweets.

In order to further improve the quality of the data, we have manually validated a subset of the data as follows: i) we have randomly selected a sample of the labeled dataset to manually validate and correct data; and ii) we selected a sample of the labelled dataset by searching for gender related words in the users' descriptions. Concerning the second task, Table 4 describes some of the words more informative about the gender. Some of these words are associated to the opposite gender when preceded by possessive determiners (e.g.: "my husband" is considered female[1], while "husband" is male). This second task may be considered as biased, since the probability of finding wrong classification is higher, but it improves the quality of the dataset.

Finally, each one of the resulting datasets were randomly partitioned into 3 subsets: *train* – includes 60% of the users and can be used to train models; ii) *development* – includes 20% of the users and can be used to train or to tune models, minimising problems, such as overfitting; iii) *test* – includes 20% of the users and can be used to assess the performance of the final models. Table 5 shows the number of tweets and users included in each one of the subsets.

4 Enriching the Datasets with Additional Information

In order to further enhance the datasets, we have added information about two new features for each user: gender recognition from profile picture, and detailed geographical information based on the last known location. The first attribute provides useful information for improving or confirming the gender classification performed previously, while the second attribute may be relevant for tackling region specific phenomena in further automatic processing.

[1] Gay and transsexual users, as profiles from companies, are not in the scope of this study.

Table 5. Split of the obtained semi-automatic gender labelled datasets.

Dataset	#tweets	train	development	test	Total
English	6.5M	39043	13015	13015	65073
Portuguese	5.8M	34625	11540	11540	57705

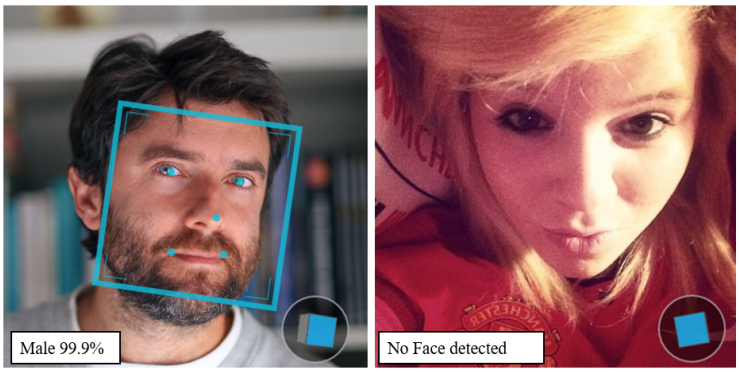

Fig. 3. Face++ gender detection examples.

4.1 Gender Based on the Profile Picture

To the best of our knowledge, the use of the gender attribute extracted from the profile picture has not been reported in previous work. However, the profile picture might contain clues regarding the gender of the user.

Face++[2]is a recent facial recognition API that is publicly available and can be used to analyse the users' profile picture. We have used it through its API to extract the gender and the corresponding confidence, and the resulting info was stored in the datasets. The API was invoked with the profile picture URL extracted from the last stored tweet of each user. Figure 3 illustrates the usage of Face++, where the first picture was correctly classified. Many of the users were correctly classified using this method, but it still presents the following limitations: i) our datasets contains data back from 2014, and some of the users have changed their profile picture in the meanwhile; ii) some of the pictures do not contain faces; iii) Face++ is sometimes unable to correctly detect the face in the picture, as exemplified in the second picture presented in the figure.

Table 6 summaries the gender data retrieved from the Face++ API, showing that 54% of the English users and 44% of the Portuguese users do not have a profile picture or have removed it since 2014. From the users with an existing profile picture, no face was detected for 36% in both datasets. In the English dataset, more male than female users have a profile picture with a face, but the opposite occurs in the Portuguese dataset. The gender information provided based on the profile picture could be combined with our previous labels in order

[2] http://www.faceplusplus.com/

Table 6. Number of users involved when Face++ was applied to guess the user gender.

	English		Portuguese	
image unavailable	31076	54%	28605	44%
no face detected	9777	17%	12995	20%
Male	9156	16%	10805	17%
Female	7857	14%	12649	19%

Table 7. Examples of geolocation information.

United States	**United Kingdom**	**Portugal**
New York, **NY**	*North East*, United Kingdom	**Lisboa**, Portugal
St. James, **NY**	Westminster, *London*	Paços de Ferreira, **Porto**
New York, US	Cardiff, **Wales**	*Vila Nova de Gaia*, Portugal
New Jersey, USA		

to enhance the classification prediction of the whole system. However, it was not used for that purpose in the scope of this paper.

4.2 Geographical Location

People may write differently according to their location, and Twitter provides geolocation within each tweet as long as the user allows it. Despite not providing additional clues about the user gender, in order to better characterise our data and provide extended usage, we have added geographical information to our datasets based on the existing metadata.

We took different approaches depending on the dataset. The English dataset contains tweets in English from more than 200 countries. Adding state or district information for each country would be almost impossible and in most cases unnecessary, since for more than 100 countries the dataset contains only a few number of users, sometimes less than 10. From the entire labelled dataset, 78% users' last geographical location was the United States and 11% the United Kingdom. For the United States' users, we added the information regarding the location's state. We extracted the last location from the users and searched for a city or state. The first and second columns of Table 7 shows examples of possible values for geolocation for the United States and United Kingdom, where bold represent states and countries. Twitter usually provides the state code for tweets geolocated in the United States (from the standard INCITS 38[3]). When the code was not found, we extracted the location and mapped it to the corresponding state code. For the United Kingdom labelled users, the distinction added was the country: Scotland, Northern Ireland, England and Wales. We extracted the last location from the users and searched for a city, a state or a country. Figure 4 shows the distribution of the labelled users in the United States and in the United Kingdom.

[3] http://geonames.usgs.gov/domestic/download_data.htm

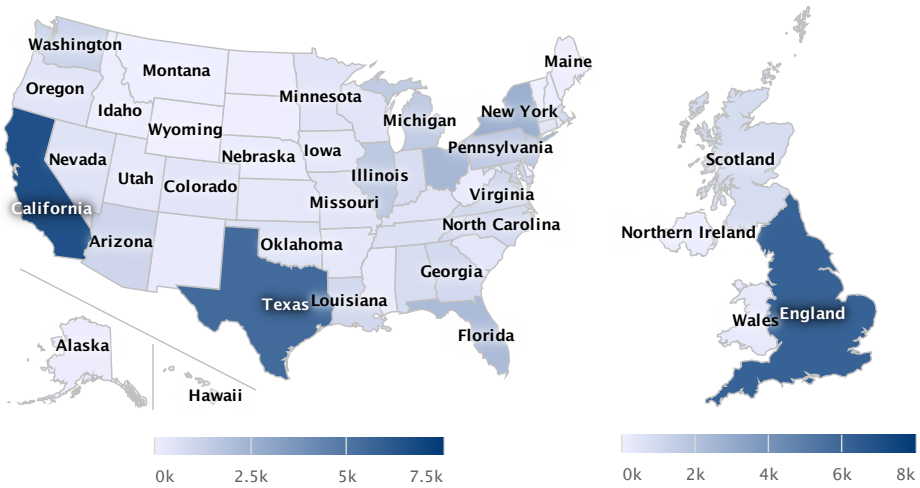

Fig. 4. Labelled users in the United States and in the United Kingdom.

Table 8. Manual validation of the automatic gender classification.

Dataset	Users					Incorrect classification					
	Total	Female		Male		Total		Female		Male	
English	3030	1883	62.2%	1147	37.9%	274	9.0%	187	68.3%	87	31.8%
Portuguese	3028	1754	57.9%	1274	42.1%	93	3.1%	76	81.7%	17	18.3%

In the Portuguese dataset, we added a feature with the district of the location. We extracted the last location from the user and searched for a city or district. After finding the cities, they were mapped to the corresponding district. The third column of Table 7 shows possible values for geolocation for the Portuguese territory, where districts are represented in bold, and cities or locations and represented in italics. In the case of the Portuguese archipelagos, we aggregated each location in its archipelago, Madeira and Azores. Finally, we added the district information to each user. Figure 5 shows the geographical distribution of Portuguese labelled users by district.

5 Data Validation

In order to independently assess the quality of the data, a manual validation was performed. About 3000 users were randomly selected from each labelled dataset, and the gender label was validated using both the Twitter profile content and the blogging sites (when available). We looked for names both in the *user name* and in the *screen name* of the profile, analysed the profile picture of the user and, if the user had blogging sites associated to their profile, we followed those URLs and cross validated the data found with their gender classification. Table 8 summaries the results obtained. We were expecting classification accuracy around 97.3%

District	Female	Male	Total
Açores	515	469	984
Aveiro	4110	2712	6822
Beja	400	292	692
Braga	2202	1427	3629
Bragança	517	323	840
Castelo Branco	1600	1324	2924
Coimbra	1715	1189	2904
Évora	440	251	691
Faro	2377	1749	4126
Guarda	66	68	134
Leiria	1384	962	2346
Lisboa	9743	8387	18130
Madeira	340	254	594
Portalegre	307	175	482
Porto	2680	1883	4563
Santarém	1179	796	1975
Setúbal	1764	1279	3043
Viana do Castelo	454	331	785
Vila Real	347	214	561
Viseu	626	431	1057

Fig. 5. Labelled Portuguese users per district and gender.

for the English dataset [23], which was the accuracy achieved for the test set of the smaller dataset, but we have detected around 9% of incorrectly classified users. In the Portuguese dataset only 3% of the users were considered incorrectly classified. That was an expected results because the Portuguese language has a construction of names with more clues to gender than English. The difference in the accuracy may be also related to the higher number of features triggered for English users, probably due to noise found in the attributes. Most of the incorrect classifications in the datasets were due to the four unavoidable reasons: i) Twitter profile was not of a person; ii) user was transsexual; iii) profile was removed and the manual validation was impossible to perform; iv) gender was incorrectly assigned. By looking at the profiles that were incorrectly classified, we noticed that female names represent a higher percentage in both datasets. In the English dataset the percentage (68%) is in accordance with variation of the sample. In the Portuguese dataset the difference is noticeable. Female users represent 82% of the errors, even though the random sample contained only 58% of female users. The overall result is very satisfactory and, to the best of our knowledge, these results are much better than any other reported non-manual gender dataset.

6 Conclusion

This paper presents an approach for creating extended gender labelled datasets of Twitter users in a semi-automatic fashion. The proposed approach was successfully applied to English and Portuguese users, and two large datasets of labeled users were created. The creation of datasets of Twitter users in commonly reported in the literature. However, most of the datasets are either not labeled with user gender or they are rather small in size. Labelled datasets reported in this paper are only surpassed in size by the work reported by Burger et al. [5], but we have employed less effort and more limited resources. The datasets obtained using the presented procedure constitute a valuable resource that can be used either for creating gender models or to perform gender dependent analyses of Twitter content. As ongoing work, we have already explored supervised and unsupervised gender classification models using the developed datasets and obtained an accuracy of 93.2% with English users and an accuracy of 96.9% with Portuguese users in our test sets.

Despite the encouraging results, the proposed approach has still several limitations that will be addressed in the near future: i) Twitter users might not use their real names and for that reason the reliability of self-declared names is uncertain (e.g.: a male user can have a female gender associated *user name*); ii) The proposed approach is not robust when facing profiles of companies and other organisations; iii) Twitter metadata might be incorrect. For example, a tweet identified by Twitter as being written in Portuguese may be written in a different language.

Acknowledgments. This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2013.

References

1. Al Zamal, F., Liu, W., Ruths, D.: Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. ICWSM 270 (2012)
2. Bamman, D., Eisenstein, J., Schnoebelen, T.: Gender in twitter: Styles, stances, and social networks. CoRR abs/1210.4567 (2012)
3. Bergsma, S., Dredze, M., Van Durme, B., Wilson, T., Yarowsky, D.: Broadly improving user classification via communication-based name and location clustering on twitter. In: HLT-NAACL. pp. 1010–1019 (2013)
4. Brogueira, G., Batista, F., Carvalho, J.P., Moniz, H.: Expanding a Database of Portuguese Tweets. In: Pereira, M.J.V., Leal, J.P., Simões, A. (eds.) 3rd Symposium on Languages, Applications and Technologies. OpenAccess Series in Informatics (OASIcs), vol. 38, pp. 275–282. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany (2014), http://drops.dagstuhl.de/opus/volltexte/2014/4576
5. 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. pp. 1301–1309. EMNLP '11, Association for Computational Linguistics, Stroudsburg, PA, USA (2011), http://dl.acm.org/citation.cfm?id=2145432.2145568

6. Ciot, M., Sonderegger, M., Ruths, D.: Gender inference of twitter users in non-english contexts. In: EMNLP. pp. 1136–1145 (2013)
7. Deitrick, W., Miller, Z., Valyou, B., Dickinson, B., Munson, T., Hu, W.: Gender identification on twitter using the modified balanced winnow (2012)
8. Fink, C., Kopecky, J., Morawski, M.: Inferring gender from the content of tweets: A region specific example. In: ICWSM (2012)
9. Halteren, H.v., Speerstra, N.: Gender recognition on dutch tweets (2014)
10. Kokkos, A., Tzouramanis, T.: A robust gender inference model for online social networks and its application to linkedin and twitter. First Monday 19(9) (2014)
11. Liu, W., Al Zamal, F., Ruths, D.: Using social media to infer gender composition of commuter populations. In: Proceedings of the when the city meets the citizen workshop, the international conference on weblogs and social media (2012)
12. Liu, W., Ruths, D.: What's in a name? using first names as features for gender inference in twitter. In: AAAI Spring Symposium: Analyzing Microtext (2013)
13. McCreadie, R., Soboroff, I., Lin, J., Macdonald, C., Ounis, I., McCullough, D.: On building a reusable twitter corpus. In: Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. pp. 1113–1114. ACM (2012)
14. Miller, Z., Dickinson, B., Hu, W.: Gender Prediction on Twitter Using Stream Algorithms with N-Gram Character Features. International Journal of Intelligence Science 2(24) (2012)
15. Nguyen, D., Trieschnigg, D., Dogruöz, A.S., Gravel, R., Theune, M., Meder, T., de Jong, F.: Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment (2014)
16. Pennacchiotti, M., Popescu, A.M.: A machine learning approach to twitter user classification. ICWSM 11, 281–288 (2011)
17. Petrović, S., Osborne, M., Lavrenko, V.: The edinburgh twitter corpus. In: Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media. pp. 25–26. WSA '10, Association for Computational Linguistics, Stroudsburg, PA, USA (2010), http://dl.acm.org/citation.cfm?id=1860667.1860680
18. 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. pp. 37–44. SMUC '10, ACM, New York, NY, USA (2010), http://doi.acm.org/10.1145/1871985.1871993
19. Ugheoke, T.O.: Detecting the Gender of a Tweet Sender. Master's thesis (2014)
20. Van Zegbroeck, E.: Predicting the Gender of Flemish Twitter Users Using an Ensemble of Classifiers. Master's thesis (2014)
21. Vicente, M., Batista, F., Carvalho, J.P.: Twitter gender classification using user unstructured information. In: Proc. of IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). Istambul, Turkey (Aug 2015), http://fuzzieee2015.org
22. Vicente, M., Carvalho, J.P., Batista, F.: Using unstructured profile information for gender classification of portuguese and english twitter users. In: IV Symposium on Languages, Applications and Tecnologies (SLATE'15). short papers, Madrid, Spain (June 2015)
23. Vicente, M., Carvalho, J., Batista, F.: Using unstructured profile information for gender classification of portuguese and english twitter users. In: Sierra-Rodríguez, J.L., Leal, J.P., Simões, A. (eds.) Languages, Applications and Technologies, Communications in Computer and Information Science, vol. 563, pp. 57–64. Springer International Publishing (12 2015), http://dx.doi.org/10.1007/978-3-319-27653-3_6