Improving Twitter Gender Classification using Multiple Classifiers *

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,
{fernando.batista,joao.carvalho}@inesc-id.pt

Abstract. The user profile information is important for many studies, but essential information, such as gender and age, is not provided when creating a Twitter account. However, clues about the user profile, such as the age and gender, behaviors, and preferences, can be extracted from other content provided by the user. The main focus of this paper is to infer the gender of the user from unstructured information, including the username, screen name, description and picture, or by the user generated content. Our experiments use an English labelled dataset containing 6.5M tweets from 65K users, and a Portuguese labelled dataset containing 5.8M tweets from 58K users. We use supervised approaches, considering four groups of features extracted from different sources: user name and screen name, user description, content of the tweets, and profile picture. A final classifier that combines the prediction of each one of the four previous partial classifiers achieves 93.2% accuracy for English and 96.9% accuracy for Portuguese data.

Keywords: gender classification, Twitter users, gender database, text mining.

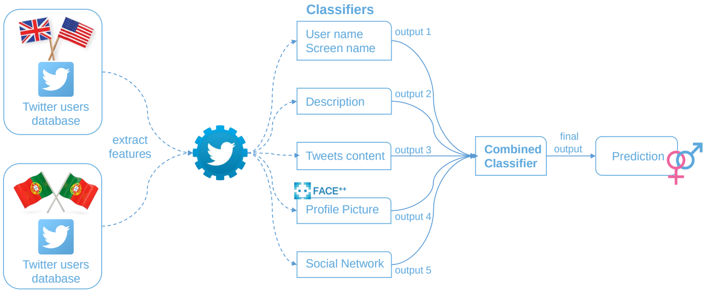
1 Introduction

Unlike other social networking services, the information provided by Twitter about a user is limited and does not specifically include relevant information. Such information is part of what can be called the user's profile, and can be relevant for a large spectra of social, demographic, and psychological studies about users' communities [6]. When creating a Twitter profile, the only required field is a user name. There are not specific fields to indicate information such as gender. Nevertheless, gender information is most of the times provided wittingly or unwittingly by the user. Knowing the gender of a Twitter user is essential for social networking studies, and useful for online marketing and opinion mining.

Our main goal is to automatically detect the gender of a Twitter user (male or female), based on features extracted from other profile information, profile picture, and the text content produced by the user. Previous research on gender detection is restricted

* Work supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) under project PTDC/IVC-ESCT/4919/2012 and reference UID/CEC/50021/2013.

Fig. 1. Combined classifier that merges the output of individual classifiers.



to features from the user generated content or from textual profile information. A relevant aspect of this study is that it involves a broader range of features, including automatic facial recognition from the profile picture. We have considered five different groups of features that were used in five separate classifiers. A final classifier, depicted in Fig. 1, combines the output of the other five classifiers in order to produce a final prediction.

This study was conducted for English and Portuguese users that produce georeferenced tweets. English is the most used language in Twitter, with 38% of the georeferenced tweets and, according to a study on 46 million georeferenced tweets [10], Portuguese is the third most used, with 6% of the georeferenced tweets. Portuguese is a morphologically rich language, contrarily to English, so interesting conclusions arise when comparing the performance achieved for both languages. Most of the previous research uses small labelled datasets, making it difficult to extract relevant performance indicators. Our study uses two large manually labelled datasets, containing 55K English and 57K Portuguese users. The proposed approach for gender detection is based on language independent features, apart from a language-specific dictionaries of first names, and can be easily extended to other Indo-European languages.

Related work

The problem of gender detection has been previously applied to Twitter. The first gender detection study applied to Twitter users was presented by [14]. The features used for gender detection were divided in four groups: network structure, communication behavior, sociolinguistic features and the content of users' postings. They achieved an accuracy of 72.3% when combining ngram-features with sociolinguistic features using the stacked Support Vector Machine based classification model.

The state-of-the-art study of [5] collected a multilingual dataset of approximately 213M tweets from 18.5M Twitter users labelled with gender. The features were restricted to word and character ngrams from tweet content and three Twitter profile fields: *description*, *screen name* and *user name*. When combining tweet text with profile information (*description*, *user name* and *screen name*), they achieved 92% of accuracy, using Balanced Winnow2 classification algorithm. [1] proposes the use of features related to the principle of homophily. This means, to infer user attributes based on the immediate neighbors' attributes using tweet content and profile information. The experiments were performed using a Support Vector Machine-based classifier and the accuracy of their prediction model was of 80.2% using neighborhood data and 79.5% when using user data only. The improvement was not considerable. [2] studies gender detection suggesting a relationship between gender and linguistic style. The experiments were performed using a logistic regression classifier and the accuracy obtained was of 88.0%. Like [1], they also study gender homophily and have the same conclusion, the homophily of a user's social network does not increase minimally the accuracy of the classifier. [9] proposes the use of neural network models for gender identification. Their limited dataset was composed of 3031 manually labelled tweets. They applied both Balanced Winnow and Modified Balanced Winnow models. Using Modified Balanced Winnow with feature selection, 53 ngram features were chosen, they achieved an accuracy of 98.5%. In a consecutive work, [13] proposes the use of stream algorithms with ngrams. They manually labelled 3000 users, keeping one tweet from each user. They use Perceptron and Naïve Bayes with character and word ngrams. They report an accuracy of 99.3% using Perceptron when tweets' length is of at least 75 characters.

Recently, some studies suggest other possible features to infer gender. [3] studied the relationship between gender, linguistic style, and social networks. They reported an accuracy of 88%. [11] studies gender classification using celebrities the user follows as features combined with tweets content features. The accuracy achieved with Support Vector Machine-based classifiers using tweets content features is of 82%. When combined with the proposed features based on the followed celebrities, the accuracy increased to 86%. [12] proposes a method to extract user attributes from the pictures posted in Twitter. They created a dataset of 10K labelled users with tweets containing visual information. Using visual classifiers with semantic content of the pictures, they achieved an accuracy of 76%. Complementing their textual classifier with visual information features, the accuracy increased from 85% to 88%.

2 Data

Experiments here described use both Portuguese and English labelled datasets from a previous study [16]. The English dataset contains 65k labelled users and the Portuguese 58k labelled users. In order to be able to train and validate the classifiers, the datasets were divided into three subsets: training, development and test.

3 Features

Twitter does not provide gender information, though the gender can be inferred from the tweets' content and the profile information. In this section, we describe the features we extract from each group of attributes. Features are distributed in the following groups: *user name* and *screen name*, *description*, tweet content, profile picture and social network.

User name and screen name. We extracted features based in self-identified names found in the *user name* and *screen name* with gender association, as proposed in our previous work [15]. In order to associate names with the corresponding gender, we used a dictionary of English names and a dictionary of Portuguese names. Both dictionaries contain *gender* and *number of occurrences* for each of the names, and focus on names that are exclusively male or female. The English names dictionary contains 8444 names. It was compiled using the list of the most used baby names from the United States Social Security Administration. The dictionary is composed of 3304 male names and 5140 female names. The Portuguese names dictionary contains 1659 names, extracted from Baptista et al. [4]. The dictionary is composed of 875 male names and 784 female names. The *user name* and *screen name* are normalized for repeated vowels (e.g.: "eriiiiiiiic"→"eric") and "leet speak" [8] (e.g.: "3ric"→"eric"). After finding one or more names in the *user name* or *screen name*, we extract the applicable features from each name by evaluating the following elements: "case", "boundaries", "separation" and "position". The final model uses 192 features.

User description. Users might provide clues of their gender in the description field. Having up to 160 characters, the description is optional. An example of user description is "I love being a mother.Enjoy every moment.". The word "mother" might be a clue to a possible female user. In order to extract useful information, we start by preprocessing the description and then we extract word unigrams, bigrams and trigrams from the preprocessed description field. We also use word count per tweet and smileys as features. Portuguese words tend to have suffixes to convey information such as gender or person and nouns inflect according to grammatical gender. For the Portuguese dataset, we also extract features related to these cases. Accordingly, if a description contains a female article followed by a word ending with the letter "a", the feature A_FEMALE_NOUN is triggered.

Content of the tweets. Features extracted from tweets' content can be divided in two groups: i) textual ngram features, like used in [5], or ii) content, style and sociolinguistic features, like emoticons, use of repeated vowels, exclamation marks or acronyms, as used in [14]. For both the textual ngram features and the style and sociolinguistic features, we only used the last 100 tweets from each labelled user. To extract textual features from tweets, we start by preprocessing the text. Retweets are ignored and the preprocessed text is used to extract unigrams, bigrams and trigrams based only on words. Though we only use word ngrams, it is advised to use character ngrams when analyzing tweets in languages like Japanese, where a word can be represented with only one character. In the study of [5], count-valued features did not improve significantly the performance. Accordingly, we also associate a boolean indicator to each feature, representing the presence or absence of the ngram in the tweet text, independently from the number of occurrences of each ngram. Besides word ngram features, we also extract content-based features, style features and sociolinguistic features that can provide gender clues. [7] suggests word-based features and function words as highly indicative of gender. We extract a group of features which include, social networks features, style features, character and word features.

Profile picture feature. Profile pictures have not been used in previous studies of gender detection of Twitter users, due to several reasons. One of the first reasons is that the profile picture is not mandatory. Also, many users tend to use profile pictures of celebrities or characters from movies and TV series. A third reason is because the picture may not be gender indicative. While the profile picture might not be good discriminating gender by itself, when combined with the other features, it might help increase significantly the accuracy of the prediction. Face++ (http://www.faceplusplus.com) is a publicly available facial recognition API that can be used to analyze the users' profile picture. We have used this tool through its API to extract the gender and the corresponding confidence. Such info was stored in our datasets. The API was invoked with the profile picture URL available on the last tweet of each user. In some cases, the API does not detect any face in the picture. 36% of the users in both datasets had no face detected. In the English dataset, more male users (34%) than female users (29%) have a profile picture with a recognizable face. In the Portuguese dataset, the opposite occurs, more female users (35%) than male users (30%) have a profile picture with a recognizable face.

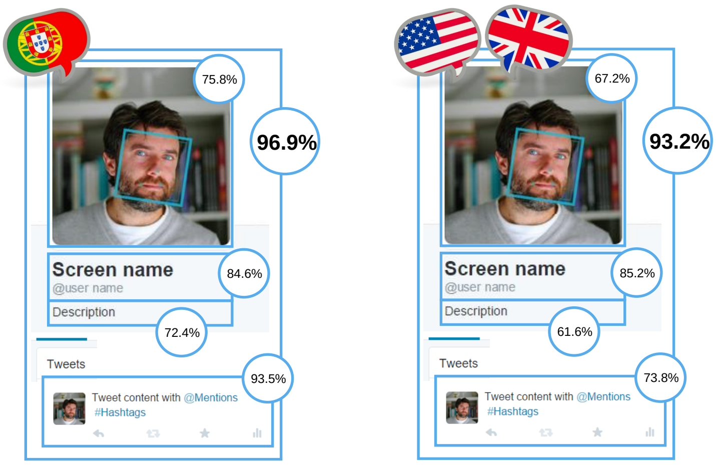
Social network features. Social network features consist in extracting the information related with the interaction between the user and other Twitter users. We extract the following attributes: Number of followers; Number of users followed; Follower-following ratio; Number of retweets; Number of replies; Number of tweets. These features alone might not be effective, but combined with the other features, could increment the global performance. We explored the extracted social network features, but we found out that these features were not indicative of gender. We observed no differences in the social network feature values between male and female. These results are consistent with the study of [14] that have analyzed users' network structure and communication behavior and observed the inability to infer gender from those attributes.

4 Experiments and results

Experiments here described use WEKA (http://www.cs.waikato.ac.nz/ml/weka) and the evaluation is performed using *Precision*, *Recall*, *F-Measure* and *Accuracy*. The combined classifier, shown in Fig. 1, receives as input the results obtained in the separate classifiers. The social network features were discarded. The separate classifiers are only used if information is available. E.g.: if a user has no description, the input from that classifier will be empty. Each classifier sends as output the confidence obtained in the classification. The values range from zero to one. If the confidence is of 100% in the class "Female," the value 1 is sent. If the confidence is of 100% in the class "Male," the value 0 is sent. If the confidence is not 100%, the values are adjusted accordingly. When the confidence received is of 0.5, we remove the input. We used an SVM to evaluate the combined classifier.

Fig. 2 summarizes the achieved accuracies per classifier for both datasets. In the Portuguese dataset we obtain 96.9% of accuracy. Only using tweets content, we already achieved an accuracy of 93.5%, but we improved the global accuracy. The experiments with the English dataset obtain an accuracy of 93.2%. With separate features, the best

Fig. 2. Classification accuracy per group of features for both datasets.



result was 85.2% using *user name* and *screen name* features. A good performance, since not all users self-assign a name in their profile information.

5 Conclusions

This study describes a method for gender detection using a combined classifier. We have used extended labelled datasets from our previous works [15, 17], partitioned into train, validation and test subsets. Instead of applying the same classifier for all features, we have grouped related features, used then in separate classifiers and then used the output of each classifier as input for the final classifier. In the Portuguese dataset, using only the tweet's text content achieves a baseline of 93.5% accuracy, but our combined classifier achieved an improved performance of 96.9% accuracy. The experiments with the English dataset achieve 93.2% accuracy. The features proposed, including the user name, screen name, profile picture and description, can be all extracted from a single tweet, except for the user text content. We successfully built two combined classifiers for gender classification of Portuguese and English users and, to our best knowledge, we provided the first study of gender detection applied to Portuguese Twitter users.

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. Bamman, D., Eisenstein, J., Schnoebelen, T.: Gender identity and lexical variation in social media. Journal of Sociolinguistics 18(2), 135–160 (2014)

4. Baptista, J., Batista, F., Mamede, N.J., Mota, C.: Npro: um novo recurso para o processamento computacional do português. In: XXI Encontro APL (Dec 2005)
5. Burger, J.D., Henderson, J., Kim, G., Zarrella, G.: Discriminating gender on twitter. In: EMNLP 2011. pp. 1301–1309. EMNLP '11, Association for Computational Linguistics, Stroudsburg, PA, USA (2011)
6. Carvalho, J.P., Pedro, V., Batista, F.: Towards intelligent mining of public social networks' influence in society. In: IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS). pp. 478 – 483. Edmonton, Canada (June 2013)
7. Cheng, N., Chandramouli, R., Subbalakshmi, K.: Author gender identification from text. Digital Investigation 8(1), 78–88 (2011)
8. Corney, M.W.: Analysing e-mail text authorship for forensic purposes. Ph.D. thesis, Queensland University of Technology (2003)
9. Deitrick, W., Miller, Z., Valyou, B., Dickinson, B., Munson, T., Hu, W.: Gender identification on twitter using the modified balanced winnow. Communications and Network 4(3) (2012)
10. Leetaru, K., Wang, S., Cao, G., Padmanabhan, A., Shook, E.: Mapping the global twitter heartbeat: The geography of twitter. First Monday 18(5) (2013)
11. Ludu, P.S.: Inferring gender of a twitter user using celebrities it follows. arXiv preprint arXiv:1405.6667 (2014)
12. Merler, M., Cao, L., Smith, J.R.: You are what you tweet... pic! gender prediction based on semantic analysis of social media images. In: Multimedia and Expo (ICME), 2015 IEEE International Conference on. pp. 1–6. IEEE (2015)
13. Miller, Z., Dickinson, B., Hu, W.: Gender prediction on twitter using stream algorithms with n-gram character features. International Journal of Intelligence Science 2(4A) (2012)
14. Rao, D., Yarowsky, D., Shreevats, A., Gupta, M.: Classifying latent user attributes in twitter. In: 2nd Int. Workshop on Search and Mining User-generated Contents. pp. 37–44. SMUC '10, ACM, New York, NY, USA (2010)
15. 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)
16. Vicente, M., Batista, F., Carvalho, J.P.: Creating extended gender labelled datasets of twitter users. In: IPMU 2016. Eindhoven, The Netherlands (June 2016)
17. Vicente, M., Carvalho, J.P., Batista, F.: Using unstructured profile information for gender classification of portuguese and english twitter users. In: SLATE'15. short papers, Madrid, Spain (June 2015)