Using Unstructured Profile Information for Gender Classification of Portuguese and English Twitter users

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

[1] INESC-ID, Lisboa, Portugal
http://www.l2f.inesc-id.pt
[2] ISCTE-IUL - Instituto Universitário de Lisboa, Lisboa, Portugal
[3] Instituto Superior Técnico, Universidade de Lisboa, Portugal

Abstract. This paper reports experiments on automatically detecting the gender of Twitter users, based on unstructured information found on their Twitter profile. A set of features previously proposed is evaluated on two datasets of English and Portuguese users, and their performance is assessed using several supervised and unsupervised approaches, including Naive Bayes variants, Logistic Regression, Support Vector Machines, Fuzzy c-Means clustering, and k-means. Results show that features perform well in both languages separately, but even best results were achieved when combining both languages. Supervised approaches reached 97.9% accuracy, but Fuzzy c-Means also proved suitable for this task achieving 96.4% accuracy.

Keywords: Twitter users; gender detection; fuzzy c-Means; supervised methods; unsupervised methods.

1 Introduction

The growth of social networks has produced massive amounts of data. This user-generated information provides clues about users' opinions, daily routines, reaction to events, among other. Twitter, with about 500 million user-generated tweets per day, provides an opportunity for social networking studies [4], and has become the subject of studies seeking to understand public opinion [7]. Unlike other social networks, a user name is the only required field when creating a Twitter profile. There are not even specific fields to indicate information such as gender or age. Nevertheless, the user profile includes optional text attributes that can be used. Previous studies support the hypothesis that users tend to choose real names more often than other forms [2] and, in fact, gender information is most of the times provided either wittingly or unwittingly, for example, in the *screen name* (e.g. "johndoe95" or "marianacruz") or in the *user name* (e.g. "John Doe the best :)" or "the macho man!!!").

The natural language processing (NLP) problem of gender detection, i.e., deciding if the author of a text is male or female, has been previously applied

to Twitter. There are basically two major ways of addressing the problem of gender detection in Twitter: 1) by looking for naming hints included in the unstructured textual profile information; 2) by analyzing the tweet contents. The first approach is *a priori* simpler, but it is highly dependent on the fact that the user must somehow hint its real name in the user name or screen name fields. On the other hand, a single tweet is enough to perform a user's gender detection. The second approach does not need such information since it looks for gender specific information (unwillingly) provided by a user when tweeting. However, it needs each user past tweeting history, and can only give good results for users that tweet a lot and produce enough text. Rao et al. [17] examined Tweets written in English, using Support Vector Machines with character ngram-features and sociolinguistic features like emoticons use or alphabetic character repetitions. They reported an accuracy of 72.3% when combining ngram-features with sociolinguistic features. The state-of-the-art study reported by Burger et al. [6] uses a large multilingual corpora, including approximately 184k users labelled with gender, 3.3 million tweets for training, and 418k tweets for testing. They used SVMs, Naive Bayes and Balanced Winnow2 with word and character N-grams as features. Using tweet texts alone they achieved the accuracy of 75.5%. When combining tweet texts with profile information (*description*, *user name* and *screen name*), they achieved 92% of accuracy. A study on Dutch users, using tokens and character n-grams, is reported by Halteren et al. [10]. Only users with significant portions of produced tweets were studied, but using SVMs and token unigrams the study reports 95.5% accuracy. In this work we try to improve automatic user gender detection in Twitter using the unstructured information found on that user profile.

Using names to detect a user gender is, a priori, a rather trivial task. All that is needed is a good dictionary of names and the will of a user to somehow provide his/her name in the profile. E.g.: the user whose user name is John Gaines, should be male. If the names appearing on the profile are not proper, e.g.: John75, JooohnGaines, or J0hn G4ines, then it is possible to recover the user name (in this case, John) using some simple text/NLP techniques. The problem is that by using such techniques, lots of noisy information might arise. In the previous example, form "John Gaines" we would obtain "John", "Aine" and "Ines". Since both Aine and Ines are female names, we would obtain a conflicting gender info. Nevertheless, using a dictionary of names and basic NLP process, the achieved accuracy is almost 89% when any form of a name is detected within the "User name" or the "Screen name" fields. It is our contention that this number can be improved by using additional features extracted from such fields.

This paper describes a set of features for gender classification proposed in our previous study [18], which rely on the user's profile unstructured textual information. The main contributions are two-fold: Firstly, we assess the performance of the features using several supervised and unsupervised methods for a Portuguese dataset, in addition to the English dataset used in our previous study. Secondly, we show that the proposed features are compatible with both languages, and that results are improved when merging both datasets. We notice

that using unsupervised methods, the increasing amount of data has positive impact on the results. The features can be used to extend gender labelled datasets for researchers.

The paper is organized as follows: Section 2 characterizes the data, describes the proposed features and describes our golden set of manually labelled data. Section 3 describes experiments and reports the corresponding results. Section 4 presents the conclusions and prospects about the future work.

2 Data and Features

Experiments performed in this paper use an English and a Portuguese dataset of Twitter users. The English dataset was extracted from one month of tweets collected during December 2014, using the Twitter *streaming/sample* API. The data has been restricted to English geolocated tweets, either from the United States or from the United Kingdom, totaling 296506 unique users. The Portuguese dataset is a subset of the data described in Brogueira et al. [5], and correspond to a database of Portuguese users, restricted by users that have tweeted during October of 2014 in Portuguese language, and geolocated in the Portuguese mainland.

2.1 Names Dictionaries

In order to automatically associate names that can be found in the user's profile with the corresponding gender, we have compiled 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, since unisex names can be classified as male or female. The English names dictionary contains about 8444 names. It was compiled using the list of the most used baby names from the United States Social Security Administration. The dictionary is currently composed of 3304 male names and 5140 female names. The Portuguese names dictionary contains 1659 names, extracted from Baptista et al. [1]. Their work is based on the extraction of names both from official institution lists and from previous corpora. The dictionary is currently composed of 875 male names and 784 female names.

2.2 Feature Extraction

Our experiments use the features proposed in a previous work [18], which are extracted with the dictionaries of names described previously. The profile information is normalized for repeated vowels (e.g.: "eriiiiiiiiic"→"eric") and "leet speak" [9] (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 elements, such as "case", "boundaries", "separation" and "position". Each feature has a minimum size threshold (i.e.: the size of the name must have at least a number of characters). Weak features have higher thresholds. If the

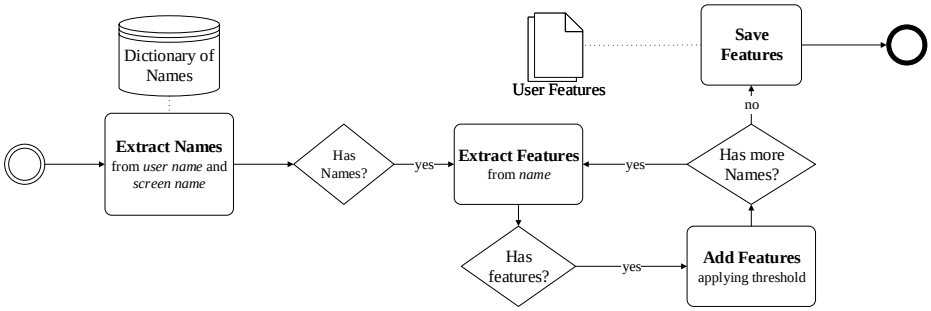

Fig. 1. Feature extraction diagram

length of the extracted name is smaller than the threshold, the feature is discarded. The final model uses 192 features. Each element increases the feature granularity. Figure 1 illustrates the feature extraction process.

Consider the screen name "john_gaines" as an example. Three names are present in the dictionary of names and are extracted: "john", "aine" and "ines". The name "aine" has no valid boundaries, since is preceeded and succeded 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 succeded by alphabetic characters. The feature for a name whith correct end boundary has a threshold of 5 and the name is discarded (e.g.: in the case of the screen name "kingjames", the name "james" whould not be discarded). Finally, the name "john" has a valid end boundary and starts at the beginning of the screen name. The feature for names with this boundary (valid end boundary) and this position (start of screen name) is 3. The name "john" is selected along with its features.

About 243522 English users (82%) and 15828 Portuguese users (58%) trigger at least one gender feature.

2.3 Labelled data

In order to perform the evaluation, we manually labelled a randomly selection of Portuguese users with gender information and used the existing labelled English dataset [18]. The corresponding gender was assigned by manually analyzing and validating users based on their user name/screen name, their profile picture and checking if associated blogging websites corresponded in gender. All users in our labelled datasets contain at least a sequence that matches a name in our dictionary of names. The English labelled dataset has 748 users: 330 male users and 418 female users. The Portuguese labelled dataset has 716 users: 249 male users and 467 female users. The majority of the users are female, which is consistent with the work of Heil et al. [11] that performed a study of correlation between name and gender, and estimates that 55% of Twitter users are female.

Table 1 shows the number of features that can be extracted from the manually labelled subset as well as statistics for the extracted names in each one of the

Table 1. Features extracted from each profile and their properties.

	English		Portuguese	
	user name	*screen name*	*user name*	*screen name*
Number of extracted features	3221	1925	1798	2404
Leet related features	291	208	17	15
Repeated vowels related features	20	48	4	122
Average Name Length (chars)	5.4	5.3	5.2	4.7
Percentage of rejected names	29%	73%	13%	16%

profile attributes. For English we observe more occurrences of features in user names (63% against 37% in screen names). The frequency of "Leet speak" is consistent with the general features distribution. As expected, repeated vowels occur more in *screen names* because they must be unique for all Twitter users, unlike *user names* that impose no restrictions to their content. For Portuguese we observe more occurrences of features in *screen names* (57% vs 47% in user names). Repeated vowels related features occur more frequently in Portuguese *screen names*. The English data reveals that names in *screen name* are more unreliable. That is due to the *screen name* being a unique string without spaces, which leads to a higher uncertainty when extracting possible names. Names in *screen name* are more unreliable in English users than in Portuguese users (73% versus 16%). A similar discrepancy can be found in Chen et al. [8], that achieved better results with Portuguese and French than with English and German, when identifying language origin of names using trigrams of letters.

3 Experiments and results

This section describes the results obtained on the English and Portuguese datasets, and the dataset containing both English and Portuguese users, when applying supervised and unsupervised approaches based on the proposed features. The supervised methods include: Multinomial Naive Bayes (MNB) [15], a variant of Naive Bayes, Logistic Regression [13], and Support Vector Machines (SVM) [16, 12]. The unsupervised methods include Fuzzy c-Means clustering (FCM) [3] and k-means [14]. The fuzzy logic toolkit for SciPy[3] was used for implementing FCM, and all the other methods were applied through Weka[4], a collection of open source machine learning algorithms and a collection of tools for data pre-processing and visualization.

While the supervised based methods use labelled data to build a model, that is not the case of unsupervised methods, which group unlabelled data into clusters. For that reason, we will first describe experiments using labelled data only, and then will extend the analysis to all the data, but restricting the experiments to unsupervised methods only. Experiments using supervised methods use the

[3] SciPy Fuzzy Logic Toolkit. https://github.com/scikit-fuzzy/scikit-fuzzy
[4] Weka version 3-6-8. http://www.cs.waikato.ac.nz/ml/weka

Table 2. Gender classification results for supervised and unsupervised methods.

	English		Portuguese		English + Portuguese	
	Accuracy	kappa	Accuracy	kappa	Accuracy	kappa
Logistic Regression	93.7%	0.87	97.6%	0.95	96.3%	0.92
Multinomial Naive Bayes	**97.2%**	**0.94**	**98.3%**	**0.96**	**97.9%**	**0.96**
Support Vector Machines	96.4%	0.93	97.8%	0.95	97.4%	0.95
kMeans clustering	67.3%		70.1%		67.8%	
Fuzzy c-Means	**96.0%**		**94.4%**		**96.4%**	

labelled data for training and used a 5-fold cross-validation. Experiments using unsupervised methods use all data for creating two different clusters, the labelled data was used for validation, and each cluster was assigned to the class with more elements from that cluster. In terms of setup, k-means was set to use the Euclidean distance, centroids are computed as a mean, and the seed was set to 10. In order to use the FCM clustering algorithm, the data has been converted into a matrix of binary values, and we have used 1000 iterations, and the Euclidean distance. All experiments consider binary features.

Results achieved with each one of the methods are summarized in Table 2. The first 3 rows show the performance for supervised methods. Results from the last two columns were achieved by combining both the English and the Portuguese labelled subsets. MNB achieved the best performance for both languages, and achieves even better performance for the merged subset of users, achieving about 98% accuracy, proving that datasets can be combined and that features are compatible with the two languages. The achieved performance suggests that the proposed features can be suitable to discriminate the user's gender for both languages. The last two rows of the table summarizes the performance for unsupervised methods. FCM obtains the correct gender for about 96.0% of the English users and about 94.4% of the Portuguese users when all the data is used. k-means achieves a much lower performance for both languages. The last column of the table shows the results when English and Portuguese data are combined. With such dataset, FCM achieves the best results so far, outperforming individual results obtained for each language.

Our proposed features compare well with the performance achieved by other state-of-the art research, despite being applied to only about 82% of English users. For example, Burger et al. [6] uses the winnow algorithm with n-grams extracted from the user's full name and obtain 89.1% accuracy for gender detection.

We have performed additional experiments in order to assess the impact of using increasing amounts of data. Figure 2 shows the impact of the amount of data on the performance of FCM, revealing that it has positive impact until reaching the 50k users. Above that threshold, the accuracy tends to remain stable, which may be due to our relatively restricted set of users.

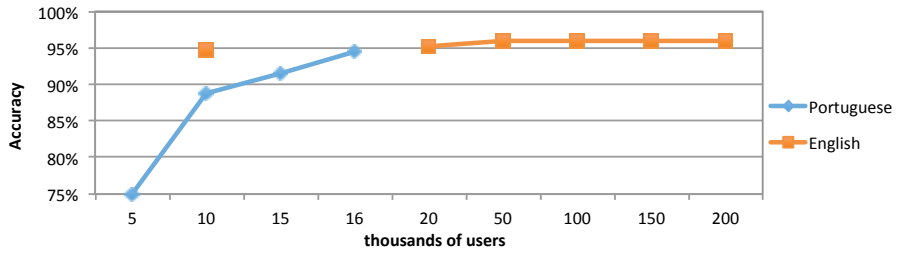

Fig. 2. Impact of the amount of data on the performance, for Portuguese and English.

4 Conclusions and future work

We have described an approach to automatically detect the gender of Twitter users, using unstructured profile information. A number of name related features is evaluated on a dataset of about 244K English users and a dataset of about 16k Portuguese users. Different supervised and unsupervised approaches are used to assess the performance of the proposed features. The proposed features proved to be good for discriminating the user's gender in Twitter, achieving about 97.9% accuracy using a supervised approaches, and about 96.4% accuracy using the unsupervised approach based on Fuzzy c-Means, which also proved to be very suitable for this task. Our features proved to be compatible between the English and Portuguese datasets of Twitter users. Experiments show that by combining datasets of English and Portuguese users, the performance can be further increased. The performance of Fuzzy c-Means significantly increased when more data was used for learning the clusters. Above 50k users, the performance stabilizes, probably to the relatively small amount of labelled data. Fuzzy c-means proved to be an excellent choice for the gender detection on Twitter since: i) it does not require labelled data, which is relevant when dealing with Twitter; ii) its performance increases as more data is provided; and iii) it achieves a performance almost similar (1.5% lower) to the best supervised method.

Future work will encompass the creation of an extended labelled dataset in a semi-automatic fashion, based on an automatic annotation provided by our proposed features. Using such labelled dataset, we will associate the textual content provided by the users with their gender and create gender models, purely based on the text contents. In addiction, we will create age models for our Twitter dataset.

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

References

1. 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)
2. Bechar-Israeli, H.: From <bonehead>to <clonehead>: Nicknames, play, and identity on internet relay chat. Computer-Mediated Communication 1(2) (1995)
3. Bezdek, J.C., Ehrlich, R., Full, W.: Fcm: The fuzzy c-means clustering algorithm. Computers and Geosciences 10(2–3), 191 – 203 (1984)
4. Brogueira, G., Batista, F., Carvalho, J.P., Moniz, H.: Portuguese geolocated tweets: An overview. In: Proceedings of the International Conference on Information Systems and Design of Communication. pp. 178–179. ISDOC '14, ACM, New York, NY, USA (2014), http://doi.acm.org/10.1145/2618168.2618200
5. Brogueira, G., Batista, F., Carvalho, J.P., Moniz, H.: Expanding a database of portuguese tweets. In: 3rd Symp. on Languages, Applications and Technologies SLATE'14. OpenAccess Series in Informatics (OASIcs), vol. 38, pp. 275–282 (2014)
6. Burger, J.D., Henderson, J., Kim, G., Zarrella, G.: Discriminating gender on twitter. In: EMNLP 2011. pp. 1301–1309. ACL (2011)
7. 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)
8. Chen, Y., You, J., Chu, M., Zhao, Y., Wang, J.: Identifying language origin of person names with n-grams of different units. In: IEEE ICASSP 2006. vol. 1, pp. I–I (May 2006)
9. Corney, M.W.: Analysing e-mail text authorship for forensic purposes. Ph.D. thesis, Queensland University of Technology (2003)
10. Halteren, H.v., Speerstra, N.: Gender recognition on dutch tweets (2014)
11. Heil, B., Piskorski, M.: New twitter research: Men follow men and nobody tweets. Harvard Business Review 1, 2009 (2009)
12. Keerthi, S., Shevade, S., Bhattacharyya, C., Murthy, K.: Improvements to platt's smo algorithm for svm classifier design. Neural Computation 13(3), 637–649 (2001)
13. Le Cessie, S., Van Houwelingen, J.C.: Ridge estimators in logistic regression. Applied statistics pp. 191–201 (1992)
14. MacQueen, J.: Some methods for classification and analysis of multivariate observations (1967), http://projecteuclid.org/euclid.bsmsp/1200512992
15. McCallum, A., Nigam, K., et al.: A comparison of event models for naive bayes text classification. In: AAAI-98 workshop on learning for text categorization. vol. 752, pp. 41–48 (1998)
16. Platt, J., et al.: Fast training of support vector machines using sequential minimal optimization. Advances in kernel methods—support vector learning 3 (1999)
17. 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. ACM (2010)
18. Vicente, M., Batista, F., Carvalho, J.P.: Twitter gender classification using user unstructured information. In: FUZZ-IEEE 2015, IEEE International Conference on Fuzzy Systems. IEEE Xplorer, Istanbul, Turkey (Accepted)