Gender Detection of Twitter Users based on Multiple Information Sources

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. Twitter provides a simple way for users to express feelings, ideas and opinions, makes the user generated content and associated metadata, available to the community, and provides easy-to-use web and application programming interfaces to access data. The user profile information is important for many studies, but essential information, such as gender and age, is not provided when accessing 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. We have performed experiments using an English labelled dataset containing 6.5M tweets from 65K users, and a Portuguese labelled dataset containing 5.8M tweets from 58K users. We have created four distinct classifiers, trained using a supervised approach, each one considering a group of features extracted from four different sources: user name and screen name, user description, content of the tweets, and profile picture. Features related with the activity, such as number of following and number of followers, were discarded, since these features were found not indicative of gender. A final classifier that combines the prediction of each one of the four previous individual classifiers achieves the best performance, corresponding to 93.2% accuracy for English and 96.9% accuracy for Portuguese data.

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

1 Introduction

With the massification of social networks, social media has become a playground for researchers. Social networks allow global communication among people, groups and organizations. The user-generated content and metadata, like geolocation, provides clues of users' behaviors, patterns and preferences. Twitter, a microblogging service, has 316 million monthly active users. On average, these users post approximately 500 million status updates, called tweets, per day. Tweets allow users to share events, daily activities, information, and to connect with friends. Twitter supports more than 35 languages and is coverage is more than global. On May 12th, 2009, astronaut Mike Massimino sent the first tweet from space. Twitter played a major role in events, like the Arab Spring [23] or The London Riots. Being an enormous source of user-generated data, Twitter has become a major tool for social networking studies. Researchers are mining Twitter generated content to extract useful information and to understand public opinion. A number of well-known tasks, including sentiment analysis, and user political orientation [12] are now being extensively applied. Twitter is also being used to practical applications such as monitoring diseases, e.g. detect flu outbreaks [14], to improving response to natural catastrophes, e.g. earthquake detection [16], or even to enhance awareness in emergency situations [33, 21].

Unlike other social networking services, the information provided by Twitter about a user is limited and does not specifically include relevant information, such as gender. 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 [9]. 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 in an unstructured form. Knowing the gender of a Twitter user is essential for social networking studies, and useful for online marketing and opinion mining.

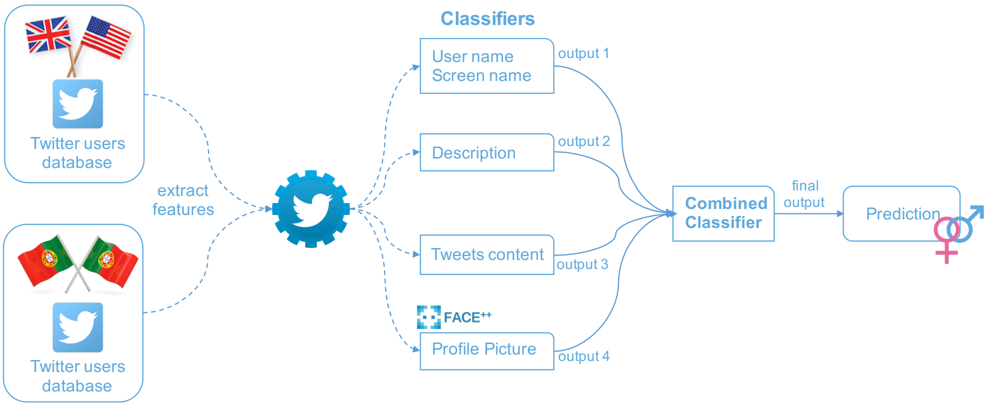


Fig. 1. Combined classifier that merges the output of individual classifiers.

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 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 four different groups of features that were used in four separate classifiers. A final classifier, depicted in Fig. 1, combines the output of the other four 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 [22], Portuguese is the third most used language in Twitter, 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, only surpassed in size by [7]. The proposed approach for gender detection is based on language independent features, except in what concerns the usage of language-specific dictionaries, and can be easily extended to other Indo-European languages.

2 Related work

A well-known Natural Language Processing (NLP) problem consists of deciding whether the author of a text is *male* or *female.* Such problem is known as gender detection or classification, and has been frequently addressed (for an overview, see e.g.: [20, 17]).

The problem of gender detection has been previously applied to Twitter. The first study was presented by Rao et. al. (2010) [27]. Their goal was to infer latent user attributes, namely: gender, age, regional origin and political orientation, and for that reason they manually annotated 500 users of each gender. The features used for gender detection were divided in four groups: network structure, communication behavior, sociolinguistic features and the content of users' postings. Both network structure features and communication behavior features had a similar distribution among genders. They reported an accuracy of 71.8% using sociolinguistic features, using ngrams they reached only an accuracy of 67.7%. They achieved an accuracy of 72.3% when combining ngram-features with sociolinguistic features using the stacked Support Vector Machine (SVM) based classification model. The study suggests Twitter sociolinguistic features to be effective for gender detection. The use of emoticons, ellipses or alphabetic character repetition indicate female users. They also observed that words following the possessive "my" have high value predicting gender.

The state-of-the-art study of [7] collected a large multilingual dataset of approximately 213M tweets from 18.5M Twitter users labeled with gender. All the users being considered have already completed a blog profile and therefore provided gender information together with the log profile. The features were restricted to word and character ngrams from tweet content and three Twitter profile fields: *description*, *screen name* and *user name.* Using tweet text alone they achieved the accuracy of 75.5%. When combining tweet text with profile information (*description*, *user name* and *screen name*), they achieved 92% of accuracy, using Balanced Winnow2 classification algorithm. Notice, that the universe of users is being restricted to users that have also created blogs, and that may be more prone to write, for example, longer tweets or tweets with more meaningful information. They further compared the automatic classification with a manual human task classification, using the Amazon Mechanical Turk (AMT). The manual human task classification achieved an accuracy of 67.3%, lower than the automatic classification. The study suggests tweet content has more gender clues than profile descriptions. *User name* proved to be the more informative field, with a performance of 84.3%, outperforming the combination of the other three fields. Also, accuracy increased when the number of tweets increased. The study supports that female users are more likely to show gender clues and update their status more often than male users. Some results were similar to those of [27]: emoticons were associated with female users while character sequences like *ht, http, htt, Googl, and Goog* were associated with male users. This study does not provide the performance of the classifiers on each different language. To further extend previous work on gender, age and political affiliation detection, [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 an SVM 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, using a 10 fold cross-validation, 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. [15] proposes the use of neural network models for gender identification.

Table 1. Datasets containing gender labelled users.

Dataset of Twitter users	#users	train	validation	test
English	65063	39043	13015	13015
Portuguese	57705	34625	11540	11540

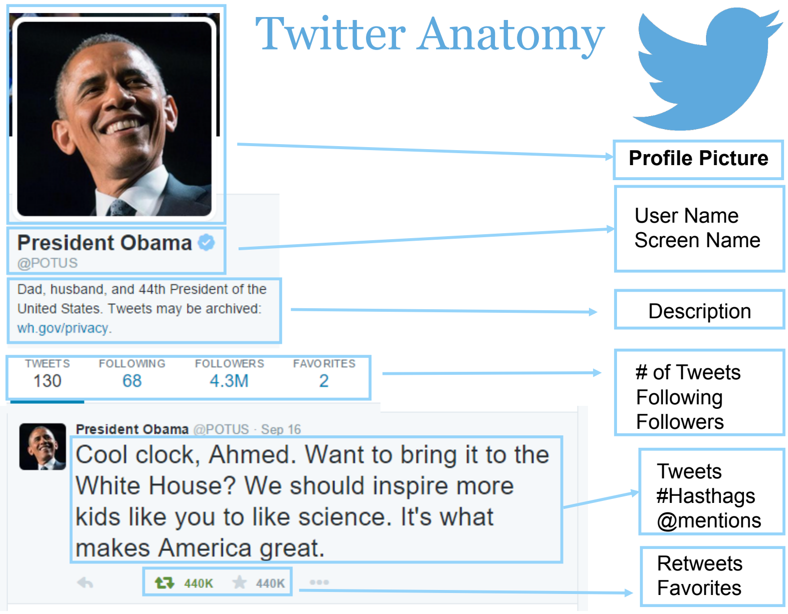
Their limited dataset was composed of 3031 manually labelled tweets, one for each user. They applied both Balanced Winnow and Modified Balanced Winnow models. In a consecutive work, [26] proposes the use of stream algorithms with ngrams. They manually labelled 3000 users, keeping one tweet from each user. They use Perceptron and Naïve Bayes with character and word ngrams. When tweets' length is of at least 75 characters, they report an accuracy of 99.3% using Perceptron.

Though the work of [7] was multilingual, the classification was global and no data was given regarding the classification of separate languages. [11] performed the first study of gender detection of non-English users. The purpose was to apply existing SVM gender classifiers to other languages and to evaluate if language-specific features could increase classification models' accuracy. They labelled users with tweets written in four different languages: Japanese, Indonesian, Turkish or French. About 1000 users per language were manually labeled. The results of French and Indonesian were comparable with the results previously obtained for English users. Turkish had a better performance and Japanese worse. After the first experiments, they created French specific features, like "je suis/*I am*" followed by an adjective. The standard classifier obtained an accuracy of 76% for French users, while the classifier with specific features for French obtained an accuracy of 83% (90% when users had tweets with "je suis"). This might not be applicable to other languages. French, like Portuguese, has gender specific nouns and adjectives.

Recently, some studies suggest other possible features to infer gender. [3] studied the relationship between gender, linguistic style, and social networks using a corpus of 14000 English Twitter users with about 9 million tweets. They reported 88% accuracy using lexical features and all user tweets. [24] studies gender classification using celebrities the user follows as features combined with tweets content features. SVM classifiers using tweets content features achieve 82% accuracy. When combined with the proposed features based on the followed celebrities, the accuracy increased to 86%. [25] 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%.

3 Data

Experiments here described use both Portuguese and English labelled datasets from a previous study [31]. This data was firstly automatically labelled based on clues provided by user profile information, using the method proposed in [31]. Later, part of the data was manually validated. The English dataset was extracted from one year of tweets collected since January until December of 2014, using the Twitter *streaming/sample* API, limited to only about 1% of the actual public tweets and restricted the data to English language and users with at least 100 tweets. The Portuguese dataset was extracted from the data described in [6], and corresponds to a database of Portuguese users, restricted by users that have tweeted in Portuguese language,


Dad, husband, and 44th President of the United States. Tweets may be archived: wh.gov/privacy.

Cool clock, Ahmed. Want to bring it to the White House? We should inspire more kids like you to like science. It's what makes America great.

Fig. 2. Anatomy of a Twitter user.

geolocated in the Portuguese mainland. We filtered the users and discarded users having less than 100 tweets. In both datasets, we retrieved only the last 100 tweets of each user. These datasets are used in the remainder of the study, unless stated otherwise. 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, as reported in Table 1. All the tweets from each user were added to the user's subset. The training subset was used to fit the parameters of the classifiers and find the optimal weights. The validation subset was used to test and tune the classifiers' parameters. Finally, the test subset was used to assess the final performance of the classifiers, avoiding biased error estimation if the validation subset was used to select the final model.

Our labelled dataset contains extended geographical information, and whereas the Portuguese dataset is restricted to the Portuguese territory, the English dataset contains tweets in English from more than 200 countries. The Portuguese dataset only contains users from Portugal. The extended geographical information contained in the dataset is the district. In the case of the Portuguese archipelagos, we aggregated each location in its archipelago, Madeira and Azores.

4 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, depicted in Fig. 2: *user name* and *screen name*, *description*, tweet content, profile picture and user activity.

4.1 User name and screen name

User name and *screen name* are valuable attributes. Online name choice has an important part in the use of social media, and users tend to choose real names more

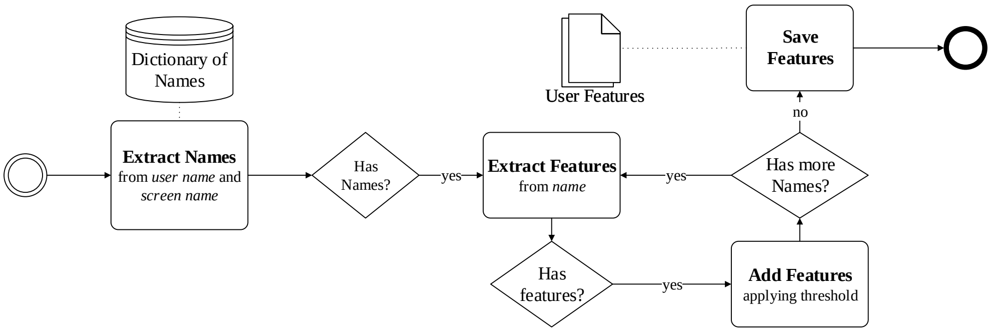


Fig. 3. Feature extraction process diagram.

often than other forms [5, 8, 28]. In the study of [28], 92% of the inquiries stated they posted real name on social media profiles. Accordingly, 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 [30]. In order to associate names with the corresponding gender, we used a dictionary of English first names and a dictionary of Portuguese first 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.

Fig. 3 illustrates the feature extraction process. The *user name* and *screen name* are normalized for repeated vowels (e.g.: "eriiiiiiiiic"→"eric") and "leet speak" [13] (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". E.g.: Consider the screen name "johnGaines" as an example. Three names are 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. The feature for a name with 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" would 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. [30] presents more details about this process. The final model uses 192 features.

4.2 User description

Users might provide clues of their gender in the description field. Having up to 160 characters, the description is optional. Table 2 lists some random descriptions from users of our labelled datasets. 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 using the following steps.

Table 2. Random Twitter user descriptions and tweets from labelled datasets.

Dataset	Gender	Description	Tweet
English	Female	I love being a mother.Enjoy every moment.	FINALLY http://t.co/NF88TgFUrq
		Sophomore • Sing • Dance • Lover • Daughter of God • Servant of the Lord	Who does that?
		19| Chill vibes only #PlayGod$™ Southern University	@KelseyAshley10 right :(I thought it was suppose to be back last month!
	Male	Southerner	First shower, then off to the barber shop to cut my hair/beard
		An ordinary person trying to do extrodinary things. Matthew 24:6	trade deadline is hockey Easter; some teams die, some rise from deadline. Hockey Christmas is the draft when everyone gets shiny new toys
Portuguese	Male	Brasileiro, casado com Ana Paula; pai de Igor Raniel e Iuri Gabriel. Pastor em Portugal. Amo Jesus, minha família e o ministério cristão.	Apenas parem lol / *Just stop lol*
		Não sei, ainda ando perdido	Bora ao cinema?? XD http://fb.me/6GNvq5YvN *let's go cinema??*
	Female	19, Moçambicana. Psicologia no ISCTE-IUL.	Ah, por favor, não se iluda. Talvez chamem você de "amor" porque esqueceram seu nome. / *Ah, don't fool yourself. Maybe they call you love because they simply forgot your name.*

– Convert all uppercase letters to lowercase letters. This allows to consider the word "Mother" the same as the word "mother";
– Replace URLs with the word URL. This way, we can use the attribute URL and can distinguish between users who share one or more URLs in the description from the ones who do not share any URL;
– Treat hashtags(#), allowing to count used hashtags and still use the word. For example "#Obama" and "obama" would both trigger the attribute *obama*, but the first example would also trigger the attribute HASHTAG;
– Replace Mentions(@) with the word "MENTION".
– Replace meta-characters. Some examples: the meta-characters "&lt;" is replaced with " LT ", "&gt;" with " GT " and "&amp;" with " & ";
– Remove special characters, punctuation and numbers;
– Extract smileys using regular expressions. E.g.: the smiley *:-)*;
– Replace accented letters with the corresponding letter without accent. E.g.: "Acção" was replaced with "accao".

After the preprocessing stage we extracted word unigrams, bigrams and trigrams. We also used 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. Some examples:

A_FEMALE_NOUN: Female articles + word ending with the letter "a".
e.g. A Geógrafa. Translated: *the geographer* (female)
A_MALE_NOUN: Male articles + word ending with the letter "o".
e.g. O Geógrafo. Translated: *the geographer* (male)
BE_FEMALE_NOUN: Auxiliary verb "Be" + word ending with the letter "a".
e.g. Sou americana. Translated: *I'm American* (female)
BE_MALE_NOUN: Auxiliary verb "Be" + word ending with the letter "o".
e.g. Sou americano. Translated: *I'm American* (male)

These features are not applicable to the English tweets, but might be useful when analyzing tweets written in Latin languages, like French, Spanish or Italian.

4.3 Content of the tweets

Features extracted from tweets' content can be divided in two groups: i) textual ngram features, like used in [7], or ii) content, style and sociolinguistic features, like emoticons, use of repeated vowels, exclamation marks or acronyms, as used in [27]. 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 [7], 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. [10] suggests word-based features and function words as highly indicative of gender. We extract a group of features which include, user activity features, style features, character and word features.

4.4 Profile picture feature

Profile pictures have not been used in previous studies of gender detection of Twitter users, due to several reasons: profile picture is not mandatory; many users tend to use profile pictures of celebrities or characters from movies and TV series; the picture may not be gender indicative; etc.. While the profile picture might not be a good gender discriminative feature 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.

4.5 User activity features

User activity 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 user activity features, but we found out that these features were not indicative of gender. We observed no differences in the user activity feature values between male and female. These results are consistent with the study of [27] that have analyzed users' network structure and communication behavior and observed the inability to infer gender from those attributes.

5 Experiments and Results

Experiments here described use WEKA (http://www.cs.waikato.ac.nz/ml/weka), an open source software with a collection of machine learning algorithms for data mining and a collection of tools for data pre-processing and visualization [18]. For most of our classification experiments, different methods have been compared, namely: Logistic Regression, Multinomial Naïve Bayes, Support Vector Machines, and C4.5 Decision Tree. The evaluation was performed using the following standard performance metrics: *Precision*, *Recall*, *F-Measure* and *Accuracy*.

5.1 Classification using user name and screen name

When the user self-assigns a name either in the *user name* or the *screen name,* the 192 features described in Section 4.1 can be used to guess the gender. The performance of such task has been previously reported in [32], but for the purpose of this study we have to consider all users, regardless of having or not a name in the profile information. If the user triggers these features, the result will be used as input in the combined classifier, otherwise it will be sent empty. The best performance for both languages was consistently achieved using Multinomial Naive Bayes.

5.2 Classification using the user description

The description field is not mandatory. For example, only 79% of the English users have a description. A number of different parameters was tested and optimized, but the best performance was achieved using word unigrams, bigrams and trigrams combined, without stemming and with stop-words. In order to reduce the number of features, we used feature selection with the evaluator Information Gain and the search algorithm Ranker (*threshold*=0). Again, Multinomial Naive Bayes achieved the best performance, with an accuracy of 61.6% for English. These results are consistent with the work of [7], where the description is the less gender indicative field. It is important to notice that the users without a description are affecting the reported performance. Some of the most strong description features of English users are similar to those presented by [7] or [29]. The top female features include *omg*, *love*, *so*, *bc*, *i love*, *cute*, *my hair*, *me*, *mom*, *hair*, *my mom*, *love you*, *i m so*, and are mostly related to sentiments or personal feelings. The top male features include *bro*, *game*, *team*, *man*, *win*, *lebron*, *my*, and are semantically related with sports or interjections, as *man* or *bro*.

5.3 Classification using tweets content

The textual features are represented using the *bag-of-words* model [19], commonly used in NLP and information retrieval (IR), where the text is represented as a set of its words, and each feature corresponds to the frequency of each word, ignoring word order or syntax. In our case, the dimension of the feature space is equal to the number of different ngrams in the last 100 tweets from all users in our test datasets.

To evaluate textual ngram features we used unigrams, bigrams, trigrams and the combination of the three. In order to test the classifiers, neither stop-words were removed nor stemming was performed. Different parameters were tested and optimized. Dimensionality reduction, TF-IDF weighting, and normalizing word frequencies increased the performance of the classifiers. We used feature selection with the evaluator Information Gain and the search algorithm Ranker (*threshold*=0). The strongest ngrams for female users are: *my hair*, *boyfriend*, *omg*, *ugh*, *cry*, *my mom*, *hair*, *cute*, *i* *love you*, *miss you*, *love you*, *i m so*, *mom*, *literally*, *seriously*, *i miss*, *so much*, *baby*, *okay*, *i hate*. The strongest ngrams for male users are: *nigga*, *man,play*, *bruh*, *game*, *games*, *the game*, *football*, *win*, *fans*, *played*, *team*, *ball*, *bro*, *beat*, *against*, *playing*, *shot*, *on the*, *go*.

Support Vector Machine using unigrams achieves the highest performance, obtaining 73.8% accuracy. Using a combination of unigrams, bigrams and trigrams, both Support Vector Machine and Logistic Regression obtain an accuracy of about 73%, but the Logistic Regression is considerably faster to build a model. We applied dimensionality reduction due to the time consumed to experiment Support Vector Machine based models. Multinomial Naive Bayes algorithms have almost a similar performance, but is more than ten times faster.

Considering we have users from more than 200 countries, we questioned if models built using only users from a specific country would increase the performance of the classifiers. For that purpose, we created a subset with users from the United States and a subset with users of the United Kingdom. The United States users represent 78% of the labelled dataset, while the United Kingdom users represent 11%. The models based on geography, and using the same parameters than before, improved the performance of almost all the methods. United Kingdom subset has only 5780 users and the performance increased slightly in Multinomial Naive Bayes and Support Vector Machine, while Logistic Regression decreased the performance. When evaluating United States subset, having 41k users, the accuracy improved in all methods. Support Vector Machine increased almost 1%, Multinomial Naive Bayes increased more than 1% and Logistic Regression increased 0.5%. Kappa, precision, recall and f-measure also increased in all methods.

As we stated previously, Portuguese words tend to have suffixes to convey information such as gender or person and nouns inflect according to grammatical gender. So, in theory it is a simpler task to predict gender using word ngrams to the Portuguese users. To evaluate textual ngram features in the Portuguese dataset, we used unigrams, trigrams, four-grams and the combination of the three. Bigrams were not used due to the lack of performance in the English users' experiments. We have replicated most of the conditions used previously for English. The best performance was achieved using SVM (93.5% accuracy) and Multinomial Naive Bayes (93.3% accuracy), outperforming the results achieved for the English dataset. The values for Kappa for SVM and Multinomial Naive Bayes are 0.851 and 0.847 respectively, indicating an excellent level of agreement.

5.4 Classification using the profile picture

To evaluate the profile picture, the Twitter profile picture is extracted and sent as parameter to the Face++ API. When a face is detected in the profile picture, we send the detected gender and confidence as input to the combined classifier. If more than one face is detected, we use the first face detected. If no face is detected, no output is sent. Even though users' profile pictures might not contain faces, or might have a picture of other person, results suggest users tend to use a picture of a matching gender. We evaluated the results in all data and in a subset of users with profile picture containing a face. The accuracy is higher in the Portuguese dataset, achieving an accuracy of 85.7% when applied to users with a face in the profile picture and 75.8% using all data. In the English dataset, the accuracy was of 76.9% in the subset of users with a face in the profile picture and 67.2% using all data. The profile picture proved to be useful for gender detection.

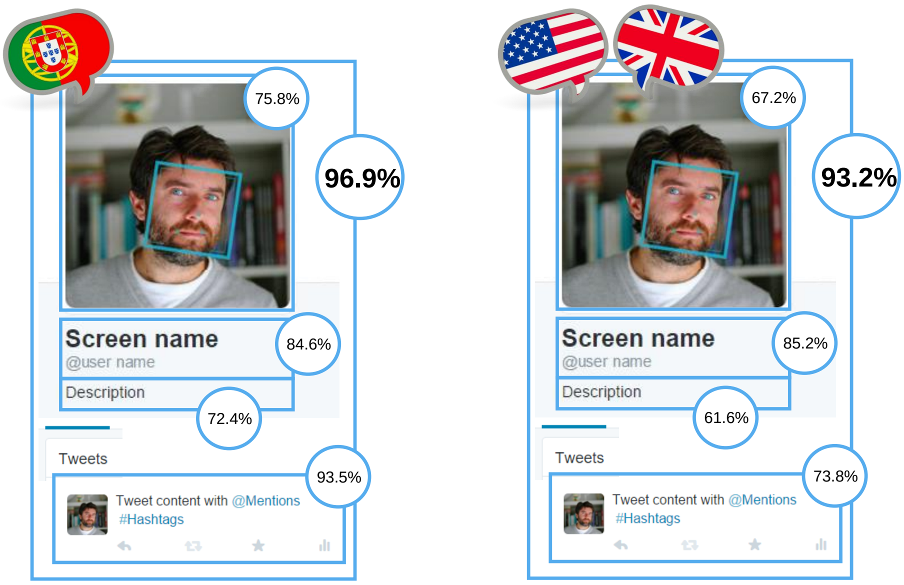


Fig. 4. Classification accuracy per group of features for both datasets.

5.5 Combined classifier

Concerning the experiments performed using individual classifiers for each group of features, the *user name* and *screen name* features reach the highest accuracy using the English dataset with 85.2%, even considering some users do not use self-assigned names in those attributes. Profile picture feature attain a lower accuracy in the English dataset, when comparing with the Portuguese dataset results. The fact that all users from the Portuguese dataset are geolocated in Portugal, while the English dataset has users from more than 200 countries, might explain the difference. In the case of the ngram features, description and tweets content, the Portuguese classifier achieves a higher accuracy by far. 93.5% of accuracy when evaluating the last 100 tweets of each user. The English classifier only achieves an accuracy of 73.8%, which is coherent with the study of [7] in a multi-language context. The description textual features were the least indicative, except for the social network features that we excluded. It must be noted that only less than 80% of the users have a description.

The combined classifier, shown in Fig. 1, receives as input the results obtained in the separate classifiers. The user activity 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 classification score, ranging from zero to one. A score near 1 indicates “Female”, while a score close to 0 indicates the “Male” class. A score of 0.5 implies removing the input. We used an SVM to evaluate the combined classifier. A number of different parameters was tested and optimized using the development set, but the best performance was achieved using the standard parameters predefined in WEKA.

Fig. 4 summarizes the achieved accuracy 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 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. With the features proposed and using the combined classifier, one tweet is enough to evaluate all features, except tweet content, namely: user name and screen name, profile picture and description features. More, using the profile picture as feature allows to evaluate user gender independently of the language used.

6 Conclusions

This study describes a method for gender detection using a combined classifier. We have used extended labelled datasets from our previous works [30, 32], partitioned into train, validation and test subsets. Instead of applying the same classifier for all features, we have grouped related features, used them 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.

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. 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. Bechar-Israeli, H.: From¡ bonehead¿ to¡ clonehead¿: Nicknames, play, and identity on internet relay chat1. Journal of Computer-Mediated Communication 1(2), 0–0 (1995)
6. Brogueira, G., Batista, F., Carvalho, J.P., Moniz, H.: Expanding a database of portuguese tweets. In: Pereira, M.J.V., Leal, J.P., Simoes, A. (eds.) SLATE'14. OpenAccess Series in Informatics (OASIcs), vol. 38, pp. 275–282. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany (2014)
7. 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)
8. Calvert, S.L., Mahler, B.A., Zehnder, S.M., Jenkins, A., Lee, M.S.: Gender differences in preadolescent children's online interactions: Symbolic modes of self-presentation and self-expression. Journal of Applied Developmental Psychology 24(6), 627–644 (2003)
9. 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)

10. Cheng, N., Chandramouli, R., Subbalakshmi, K.: Author gender identification from text. Digital Investigation 8(1), 78–88 (2011)
11. Ciot, M., Sonderegger, M., Ruths, D.: Gender inference of twitter users in non-english contexts. In: EMNLP. pp. 1136–1145 (2013)
12. Conover, M., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F., Flammini, A.: Political polarization on twitter. In: ICWSM (2011)
13. Corney, M.W.: Analysing e-mail text authorship for forensic purposes. Ph.D. thesis, Queensland University of Technology (2003)
14. Culotta, A.: Detecting influenza outbreaks by analyzing twitter messages. arXiv preprint arXiv:1007.4748 (2010)
15. 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)
16. Earle, P., Guy, M., Buckmaster, R., Ostrum, C., Horvath, S., Vaughan, A.: Omg earthquake! can twitter improve earthquake response? Seismological Research Letters 81(2), 246–251 (2010)
17. Eckert, P., McConnell-Ginet, S.: Language and gender. Cambridge University Press (2013)
18. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: Weka data mining software: an update. SIGKDD Explor. Newsl. 11(1), 10–18 (2009)
19. Harris, Z.S.: Distributional structure. Word (1954)
20. Holmes, J., Meyerhoff, M.: The handbook of language and gender, vol. 25. John Wiley & Sons (2008)
21. Imran, M., Castillo, C., Diaz, F., Vieweg, S.: Processing social media messages in mass emergency: A survey. ACM Computing Surveys (CSUR) 47(4), 67 (2015)
22. Leetaru, K., Wang, S., Cao, G., Padmanabhan, A., Shook, E.: Mapping the global twitter heartbeat: The geography of twitter. First Monday 18(5) (2013)
23. Lotan, G., Graeff, E., Ananny, M., Gaffney, D., Pearce, I., et al.: The arab spring—the revolutions were tweeted: Information flows during the 2011 tunisian and egyptian revolutions. International journal of communication 5, 31 (2011)
24. Ludu, P.S.: Inferring gender of a twitter user using celebrities it follows. arXiv preprint arXiv:1405.6667 (2014)
25. 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)
26. Miller, Z., Dickinson, B., Hu, W.: Gender prediction on twitter using stream algorithms with n-gram character features. Int. Journal of Intelligence Science 2(4A) (2012)
27. 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)
28. Stopczynski, A., Sekara, V., Sapiezynski, P., Cuttone, A., Madsen, M.M., Larsen, J.E., Lehmann, S.: Measuring large-scale social networks with high resolution. PloS one 9(4), e95978 (2014)
29. Van Zegbroeck, E.: Predicting the gender of flemish twitter users using an ensemble of classifiers (2014)
30. 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)
31. Vicente, M., Batista, F., Carvalho, J.P.: Creating extended gender labelled datasets of twitter users. In: IPMU 2016. Eindhoven, The Netherlands (June 2016)
32. 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)
33. Vieweg, S., Hughes, A.L., Starbird, K., Palen, L.: Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In: Proceedings of the SIGCHI conference on human factors in computing systems. pp. 1079–1088. ACM (2010)