PLOS ONE

A Combined Approach to Twitter Gender Classification Using Multimodal Information - A Case Study of Portuguese and English Users

--Manuscript Draft--

Manuscript Number:	
Article Type:	Research Article
Full Title:	A Combined Approach to Twitter Gender Classification Using Multimodal Information - A Case Study of Portuguese and English Users
Short Title:	A Combined Approach to Twitter Gender Classification Using Multimodal Information
Corresponding Author:	Joao Paulo Carvalho INESC-ID / Instituto Superior Tecnico, Universidade de Lisboa Lisboa, PORTUGAL
Keywords:	Twitter; gender detection; multimodal classification; supervised methods; unsupervised methods.
Abstract:	Existing social networks provide easy means for people to communicate and express their feelings. Twitter, in particular, is nowadays being extensively used, and has become a relevant source of information for many studies in different domains. Twitter 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 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 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 predition of each one of the four previous partial classifiers achieves 93.2% accuracy for English and 96.9% accuracy for Portuguese data. The proposed methodology is language independent, and can easily ported to other Indo-European languages.
Order of Authors:	Marco Vicente
	Joao Paulo Carvalho
	Fernando Batista
Opposed Reviewers:	
Additional Information:	
Question	**Response**
Financial Disclosure Please describe all sources of funding that have supported your work. A complete funding statement should do the following: Include **grant numbers and the URLs** of any funder's website. Use the full name, not acronyms, of funding institutions, and use initials to identify authors who	This work was supported by national funds through Fundação para a Ciência e a 668 Tecnologia (FCT) under project PTDC/IVC-ESCT/4919/2012 and funds with reference 669 UID/CEC/50021/2013.

received the funding. **Describe the role** of any sponsors or funders in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. If they had no role in any of the above, include this sentence at the end of your statement: "*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*" If the study was **unfunded**, provide a statement that clearly indicates this, for example: "*The author(s) received no specific funding for this work.*" * typeset	
Competing Interests You are responsible for recognizing and disclosing on behalf of all authors any competing interest that could be perceived to bias their work, acknowledging all financial support and any other relevant financial or non-financial competing interests. Do any authors of this manuscript have competing interests (as described in the PLOS Policy on Declaration and Evaluation of Competing Interests)? **If yes**, please provide details about any and all competing interests in the box below. Your response should begin with this statement: *I have read the journal's policy and the authors of this manuscript have the following competing interests:* **If no** authors have any competing interests to declare, please enter this statement in the box: "*The authors have declared that no competing interests exist.*" * typeset	The authors have declared that no competing interests exists
Ethics Statement You must provide an ethics statement if	N/A

your study involved human participants, specimens or tissue samples, or vertebrate animals, embryos or tissues. All information entered here should **also be included in the Methods section** of your manuscript. Please write "N/A" if your study does not require an ethics statement. **Human Subject Research (involved human participants and/or tissue)** All research involving human participants must have been approved by the authors' Institutional Review Board (IRB) or an equivalent committee, and all clinical investigation must have been conducted according to the principles expressed in the Declaration of Helsinki. Informed consent, written or oral, should also have been obtained from the participants. If no consent was given, the reason must be explained (e.g. the data were analyzed anonymously) and reported. The form of consent (written/oral), or reason for lack of consent, should be indicated in the Methods section of your manuscript. Please enter the name of the IRB or Ethics Committee that approved this study in the space below. Include the approval number and/or a statement indicating approval of this research. **Animal Research (involved vertebrate animals, embryos or tissues)** All animal work must have been conducted according to relevant national and international guidelines. If your study involved non-human primates, you must provide details regarding animal welfare and steps taken to ameliorate suffering; this is in accordance with the recommendations of the Weatherall report, "The use of non-human primates in research." The relevant guidelines followed and the committee that approved the study should be identified in the ethics statement. If anesthesia, euthanasia or any kind of animal sacrifice is part of the study, please include briefly in your statement which substances and/or methods were applied.	

Please enter the name of your Institutional Animal Care and Use Committee (IACUC) or other relevant ethics board, and indicate whether they approved this research or granted a formal waiver of ethical approval. Also include an approval number if one was obtained. **Field Permit** Please indicate the name of the institution or the relevant body that granted permission.	
Data Availability PLOS journals require authors to make all data underlying the findings described in their manuscript fully available, without restriction and from the time of publication, with only rare exceptions to address legal and ethical concerns (see the PLOS Data Policy and FAQ for further details). When submitting a manuscript, authors must provide a Data Availability Statement that describes where the data underlying their manuscript can be found. Your answers to the following constitute your statement about data availability and will be included with the article in the event of publication. **Please note that simply stating 'data available on request from the author' is not acceptable.** ***If, however, your data are only available upon request from the author(s), you must answer "No" to the first question below, and explain your exceptional situation in the text box provided.*** Do the authors confirm that all data underlying the findings described in their manuscript are fully available without restriction?	Yes - all data are fully available without restriction
Please describe where your data may be found, writing in full sentences. **Your answers should be entered into the box below and will be published in the form you provide them, if your manuscript is accepted.** If you are copying our sample text below, please ensure you replace any instances of XXX with the appropriate details. If your data are all contained within the	Data can be made available upon request to the authors, and will be made available on a public repository if the paper is accepted.

paper and/or Supporting Information files, please state this in your answer below. For example, “All relevant data are within the paper and its Supporting Information files.” If your data are held or will be held in a public repository, include URLs, accession numbers or DOIs. For example, “All XXX files are available from the XXX database (accession number(s) XXX, XXX)." If this information will only be available after acceptance, please indicate this by ticking the box below. If neither of these applies but you are able to provide details of access elsewhere, with or without limitations, please do so in the box below. For example: “Data are available from the XXX Institutional Data Access / Ethics Committee for researchers who meet the criteria for access to confidential data.” “Data are from the XXX study whose authors may be contacted at XXX.” * typeset	
Additional data availability information:	Tick here if the URLs/accession numbers/DOIs will be available only after acceptance of the manuscript for publication so that we can ensure their inclusion before publication.

Lisboa, November 15th, 2015

Dear Editor-in-Chief

Please find attached the paper entitled A Combined Approach to Twitter Gender Classification Using Multimodal Information - A Case Study of Portuguese and English Users ", that we are submitting for possible publication in PLOSONE.
The main focus of the paper is to infer the gender of a Twitter user from unstructured information, including the username, screen name, description and picture, or by the user generated content. Twitter gender detection is a relevant issue since Twitter profiles do not have a gender field. We obtain very interesting results using an original approach that largely improve previous studies.
We believe that the work is of interest for a large community and within the scope of this journal.

Looking forward to receiving your news,

Sincerely,

João Paulo Carvalho
Fernando Batista
Marco Vicente

Mailing address:
INESC-ID
Rua Alves Redol, 9, 1000-029 Lisboa, Portugal
E-mail: joao.carvalho@inesc-id.pt
Tel.: +351962148272

A Combined Approach to Twitter Gender Classification Using Multimodal Information - A Case Study of Portuguese and English Users

Marco Vicente[1,2], Fernando Batista[1,2], Joao Paulo Carvalho[1,3]

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

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

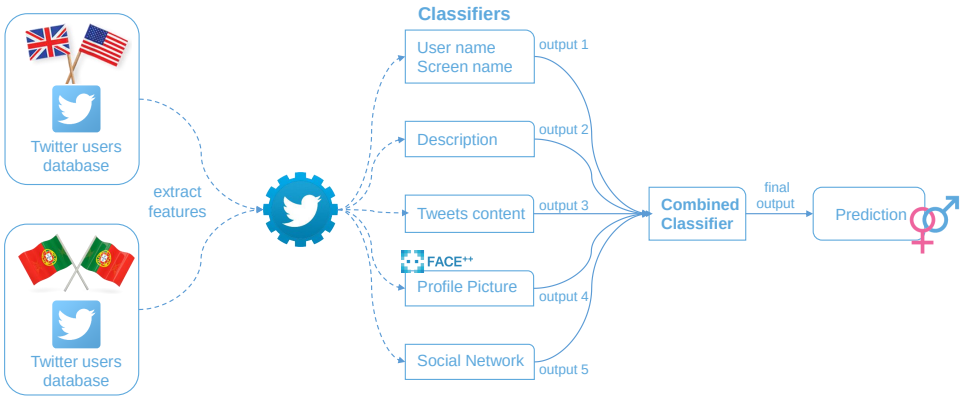
Abstract

Existing social networks provide easy means for people to communicate and express their feelings. Twitter, in particular, is nowadays being extensively used, and has become a relevant source of information for many studies in different domains. Twitter 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 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 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 predition of each one of the four previous partial classifiers achieves 93.2% accuracy for English and 96.9% accuracy for Portuguese data. The proposed methodology is language independent, and can easily ported to other Indo-European languages.

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, provide 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, connect with friends. Twitter supports more than 35 languages and has a truly more than global coverage. On 12 May 2009, astronaut Mike Massimino

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



sent the first tweet from space. Twitter has been influential in social events, like the Arab Spring [1]. 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, user political orientation [2] are now being extensively applied. Twitter is also being used to practical applications, such as to monitor diseases, e.g. detect flu outbreaks [3], to improve response to natural catastrophes, e.g. detect earthquakes [4], or even to enhance awareness in emergency situations [5, 6].

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 [7]. 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, but it is available in an unstructured form. Knowing the gender of a Twitter user is essential for social networking studies and useful for online marketing. Opinion mining, like sentiment analysis, need users' attributes, like gender, location and age. Twitter recently introduced a birthday field in the profile, but the users' gender can only be inferred. In a gender related marketing campaign, for example to an "after-shave", the ability to target male users is useful, because female users are less likely to be interested in such a campaign. The gender information allows advertising to be effective and social studies to be more accurate.

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 and 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, allowing to assess the contribution of each group of features. 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 with 38% of the georeferenced

tweets but, according to a study on 46 million georeferenced tweets [8], 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 [9]. 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 in other Indo-European languages.

Related work

A well-known Natural Language Processing (NLP) problem consists of deciding whether the author of a text is *male* or *female*. Such a problem is known as gender detection or classification, and is often addressed [10–19].

The study of the relation between gender and language usage is extensive (for an overview, see e.g.: [20, 21]). Research has been published which supports the hypothesis that analyzing linguistic features associated with male or female use of language, it is possible to detect users' gender [22–24]. [10], using automated text categorization techniques, report gender detection with approximately 80% accuracy using function words and parts of speech to infer the gender. In a later research [25], two of the authors of the former study (Schler and Koppel), assembled a large corpus of blogs (Blog Authorship Corpus) labelled for a variety of demographic attributes, including author-provided indication of gender, with over 71000 blogs. This corpus was later used by [14] to discuss and experiment more complex variants for authorship attribution, including gender detection. They report an accuracy of 72.0% using word classes derived from systemic functional linguistics and 75.1% accuracy using character ngrams. When combining style features with content features, they achieved an overall accuracy of 76.1%. This corpus was used by [13]. They improved the overall accuracy to 89.2%, using average sentence length, usage of slang and usage of non-dictionary words. [16] studied gender identification using two large text datasets: a large collection of Reuters News stories, and Enron email dataset, containing emails from about 150 users, mostly senior management of Enron. They applied three different supervised classification techniques: support vector machine, Bayesian logistic regression and AdaBoost decision tree. Using linguistic and stylometric features, obtained an accuracy of 85.1% on gender prediction using Support Vector Machine. [17] used a corpus of about 1.5M Flemish Dutch Netlog posts for gender classification. Netlog is a Belgian online social networking platform (http://nl.netlog.com/). The corpus was labelled with the age, gender and location of the authors. The features selected were word unigrams, bigrams, and trigrams, and also character bigrams, trigrams and tetra grams. They achieved an accuracy of 88.8% using a SVM classification model. [26] studied gender classification of web blogs, using part-of-speech tagging and language model features. They used several classification models based on decision trees, support vector machines and lazy-learning algorithms. Random forest classification model outperformed other models, achieving an accuracy of 70.5%. [19] presents an overview of the existing research analyzing the differences between genders in the usage of microblogs.

The problem of gender detection 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.

The first gender detection study applied to Twitter users was presented by [27]. Their goal was to infer latent user attributes, namely: gender, age, regional origin and political orientation. 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 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 [9] collected a large multilingual dataset labelled with gender. While [27] manually annotated 1000 English users, [9] created a corpus of approximately 213M tweets from 18.5M Twitter users labelled with gender. The most representative languages in the corpus are English (67%), Portuguese (14%) and Spanish (6%). The features were restricted to word and character ngrams from tweet content and three Twitter profile fields: *description*, *screen name* and *user name*. The features were boolean, representing the presence or absence of the ngram, not counting the number of occurrences of the same ngram for each user. The features appearing in less than three users were ignored. Results presented are global, and the accuracy for each language is not revealed. The experiments were performed using Support Vector Machines, Naive Bayes and Balanced Winnow2 machine learning algorithms to build gender classification models. 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. 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, [28] propose 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. Homophily suggests users connected with similar users occurs at a higher rate than among different users and previous studies suggest homophily establishes similarity between connected users [29]. 400 users were manually labeled using the self-reported first names of their user profile. The name had to be one of the 100 most common names for babies born in the United States, as reported by the U.S. Social Security Administration (technique first proposed by [30]). The last 1000 tweets from both the labelled users and all followed users were collected. The features

selected from user and neighborhood data were k-top words, k-top stems, k-top bigrams and trigrams, k-top hastags, frequency statistics, retweeting tendency and neighborhood size. The experiments were performed using a Support Vector Machine-based classifier, using a 10-fold cross-validation. In the case of gender, 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, unlike age and political affiliation, where the proposed features improved the accuracy up to 35%. In a posterior study [31], three of the four elements (Liu, Al Zamal and Ruths) applied the same gender inference algorithm to Toronto's commuting population. The objective was to infer the gender of Toronto commuting users of three modes of transportation: cars, public transportation and bicycles. They identified popular accounts dedicated to broadcasting news about Toronto's traffic, public transportation and cycling. For each Twitter user following these accounts, the most recent 1000 tweets were extracted. In each category, 4000 users were manually labeled using both user profile information and user tweets content. The proposed model achieved a gender prediction accuracy of 84.7% for public transportation, 81.0% for cars and 73.8% for bicycles.

[32] study gender detection suggesting a relationship between gender and linguistic style. They also investigate social network connection features. Using the Twitter streaming API, they collected American English users, by requiring from all filtered accounts the use of at least 50 of the 1000 most common words in the US English. The 1000 words are not specified in the study. They manually labelled authors using the census information from the US Social Security administration. Users' first names were taken into account to assign gender to Twitter authors and no data validation is mentioned. The resulting dataset contained approximately 14.4k users and 9.2M tweets. The lexical features were word unigrams. The experiments were performed using a logistic regression classifier, using a 10 fold cross-validation. The accuracy obtained was of 88.0%. Like [28], 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.

[33] propose 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, [34] proposes the use of stream algorithms with ngrams. They manually labelled 3000 users, keeping one tweet from each user. They use Perceptron and Naive Bayes with character and word ngrams. They report an accuracy of 99.3% using Perceptron when tweets' length is of at least 75 words.

[35] present a region-specific study, focusing on Nigerian Twitter users. They label users based on the tweets' geolocation to create their dataset. Their experiments use Support Vector Machine with a linear kernel implementation [36], based on word unigrams, hashtags and Linguistic Inquiry and Word Count, or LIWC [37]. They report an accuracy of 81% when using unigrams as features.

While the previous studies focused on tweets' content alone, [38] study the connection between gender and the self-reported first name. They add name features to tweets ngram features. Using an Support Vector Machine classifier, they improved the accuracy from a baseline of 83%, using only ngrams, to 87%, using also first name features. [39] studies gender detection based on users' preferences and location. They classify using distributed k-Means clustering and Support Vector Machine with character ngram and token features. Token features are booleans for first names, having 1 if the name is present in the profile and 0 if not. They report an accuracy of 90% when combining cluster features with ngrams and token features.

Though the work of [9] was multilingual, the classification was global and no data

Table 1. Datasets containing gender labelled users.

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

was given regarding the classification of separate languages. [40] performed the first study of gender detection of non-English users. The purpose was to apply existing Support Vector Machine 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" 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.

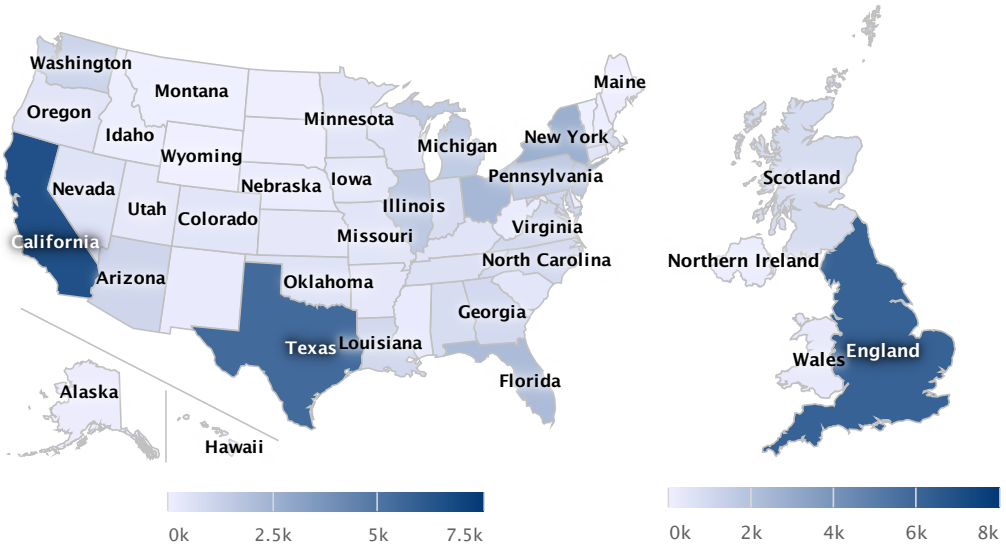
Recently, some studies suggest other possible features to infer gender. [41] studied the relationship between gender, linguistic style, and social networks using a corpus of 14000 English Twitter users with about 9 million tweets. They reported an accuracy of 88% using lexical features, when using all user tweets. [42] 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%. [43] propose 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%.

Data

Experiments here described use both Portuguese and English labelled datasets from a previous study [44]. This data was firstly automatically labelled based on clues provided by user profile information, using the method proposed in [45]. 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 [46], and corresponds to a database of Portuguese users, restricted by users that have tweeted in Portuguese language, 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.

A sample of both labelled datasets containing 3000 users and associated gender, is available in the supporting information accompanying this paper (EN-Dataset and PT-Dataset). In order to be able to train and validate the classifiers, the datasets were divided into three subsets: training, development and test, with the sizes shown 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.

Figure 2. Labelled users in the US and UK, respectively.



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. From the entire labelled dataset, 78% of users' last geographical location was the United States and 11% the United Kingdom. Regarding United States users, the state from the last geolocated tweet and from the United Kingdom users the country: Scotland, Northern Ireland, England and Wales. Fig. 2shows the distribution of the labelled users in the United States (left), and in the United Kingdom (right).

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. Fig. 3 shows a geographical distribution of the Portuguese labelled users.

Features

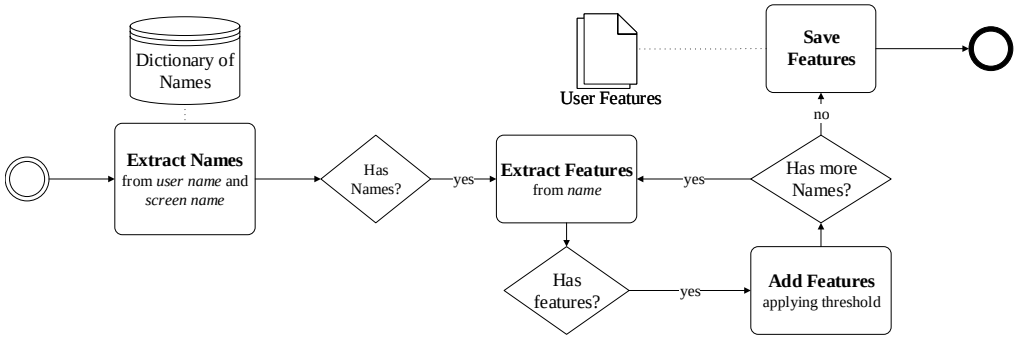
Twitter does not provide gender information, though the gender can be inferred from the tweets' content and the profile information. In this section, we describe the features we extract from each group of attributes. Features are distributed in the following groups: *user name* and *screen name*, *description*, tweet content, profile picture and social network. Fig. 4 shows different attributes that may provide clues to infer the user gender.

All feature extraction algorithms were implemented using Python 3.4. Data preprocessing and transformation routines were also developed in Python with the support of the NLTK (Natural Language Toolkit) 3.0 package (http://www.nltk.org/). NLTK provides a collection of NLP modules.

Figure 3. Labelled users in the Portuguese territory, grouped by district.

Figure 4. Anatomy of a Twitter user.

Figure 5. Feature extraction process diagram.



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 often than other forms [47–49]. In the study of [49], 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 [45]. 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. [50]. The dictionary is composed of 875 male names and 784 female names.

Fig. 5 illustrates the feature extraction process. The *user name* and *screen name* are normalized for repeated vowels (e.g.: "eriiiiiiiiic"→"eric") and "leet speak" [51] (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. The final model uses 192 features.

From the dataset of English users, 82% triggered at least one feature and from the Portuguese dataset, 58% triggered features.

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.

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
		Não sei, ainda ando perdido	Bora ao cinema?? XD http://fb.me/6GNvq5YvN
	Female	19, Moçambicana. Psicologia no ISCTE-IUL.	Ah, por favor, não se iluda. Talvez chamem você de “amor” porque esqueceram seu nome.

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

In one of the examples, the 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 preprocess the description information with the following steps:

- 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;
- Replace Hashtags(#) with the word “HASHTAG”. This allows to count used hastags and still use the word. As 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 “<” is replaced with “ LT ”, “>” with “ GT ” and “&” 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”.

Figure 6. Most used words by English female and male users, respectively.

After the preprocessing, 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 articles 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.

Content of the tweets

Features extracted from tweets' content can be divided in two groups: i) textual ngram features, like used in [9], 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. Table 2 lists some random tweets from our labelled datasets.

Textual ngram features. To extract textual features from tweets, we previously preprocess the text as described in previous Subsection of "Description" features. 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 [9], 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. Fig. 6 show the

Table 3. Examples of style and sociolinguistic features.

Social Network Features	
Instagram, facebook, snapchat, tumblr, blogspot, wordpress, linkedin, pinterest, flickr, hi5, myspace, messenger	
Style Features	
Smileys	example: **:-)**
Repeated letters	example: nooooooooooooo
Acronyms	example: LOL, ROLF
Number of exclamation marks, question marks, multiple exclamation or question marks	
Character Features	
Number of characters	
Number of letters [a-z]	
Number of digits [0-9]	
Number of uppercase letters	
Number of special characters	
Word Features	
Number of words	
Average length of words	
Number of different words	
Number of words longer than 6 characters	

most used words of female and male English users respectively. From the most used 1000 words, almost 70% of the words have a length of 5 or less characters. 68.6% from female users and 68.5% for male users.

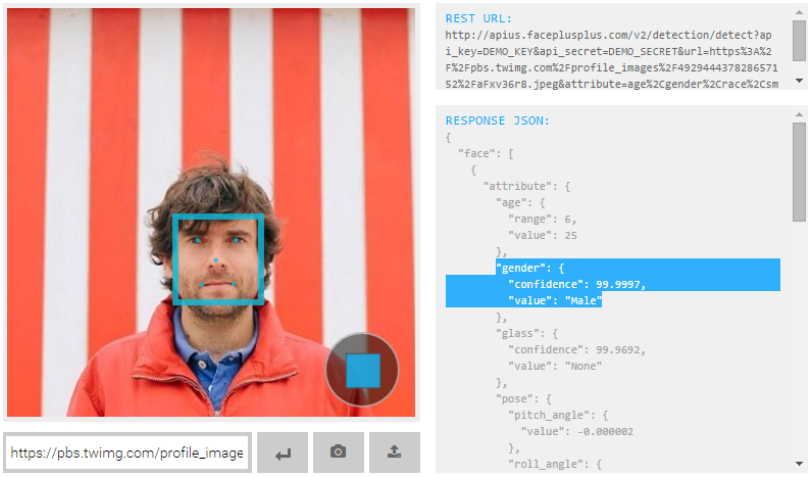
Style and sociolinguistic features. Besides word ngram features, we also extract content-based features, style features and sociolinguistic features that can provide gender clues. [16] suggest word-based features and function words as highly indicative of gender. We extract a group of features which include, social networks features, style features, character and word features. Table 3 lists some of our features that were extracted by using regular expressions.

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 use 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. Fig. 7 illustrates the usage of Face++, where the picture was correctly classified.

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.

Figure 7. Face++ gender detection.



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 [27] that have analyzed users' network structure and communication behavior and observed the inability to infer gender from those attributes.

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 [52]. To perform our experiences in WEKA, we created datasets in the ARFF (Attribute-Relation File Format) file format, thus easily allowing to use the same data to apply different Machine Learning schemes.

The evaluation is performed using four standard evaluation metrics: *Precision*, *Recall*, *F-Measure* and *Accuracy*, defined as follows:

$$Precision = \frac{\#TP}{\#TP + \#FP}, \quad (1)$$

$$Recall = \frac{\#TP}{\#TP + \#FN}, \tag{2}$$

$$F - Measure = 2 \times \frac{Precision \times Recall}{Precision + Recall}, \tag{3}$$

$$Accuracy = \frac{\#TP + \#TN}{\#TP + \#TN + \#FP + FN}, \tag{4}$$

In the equations 1, 2, 3 and 4, TP (true positive) corresponds to a user being correctly identified as belonging to a given gender; FP (false positive) corresponds to a user not belonging to a given gender and incorrectly classified with that gender; TN (true negative) corresponds to a user not belonging to a given gender and correctly identified as so; FN (False Negative) corresponds to a user belonging to a given gender, has been incorrectly identified as so.

Data representation

In order to predict gender, the relevant sources of information are the text contained in each tweet and the user profile information. We already described the features extracted and the preprocessing applied. However, some of the features are composed of text, and text is an unstructured form of data. Classifiers cannot process unstructured information [53]. For that reason, our information must be converted into vectors for each classifier, representing the user attributes. Each classifier receives a different vector representation as each classifier receives different attributes. A vector ${}^n\nu = (x_1, x_2, x_3, ..., x_n)$ has as many elements as features. Element x_1 corresponds to a feature and has zero if the feature does not occur or one if the feature occurs at least once. In the case of the social network features and some of the style and sociolinguistic features, the element is filled with the number of occurrences. E.g.: Feature "number of uppercase letters" will be filled with the number of times an uppercase letter occurs in the tweets of the user.

The textual ngram features will be represented using the *bag-of-words* model [54]. This model is used in NLP and information retrieval (IR). The text is represented as a set of its words, 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. The following example illustrates this model of representation:

```
Fav if you love Naruto!
I LOVE YOU
love the void
```

Using these three tweets, we create a dictionary *{fav, if ,you, love, naruto, i, the, void}*. The tweets can be represented as a matrix containing as much elements as the number of distinct words and with three rows, corresponding to each tweet.

```
array([
  [1,1,1,1,1,0,0,0]
  [0,0,1,1,0,1,0,0]
  [0,0,0,1,0,0,1,1]
])
```

Table 4. Gender classification results for user name and screen name features.

	English		Portuguese	
	Accuracy	Kappa	Accuracy	Kappa
Baseline	54.3%		60.8%	
Logistic Regression	81.4%	0.631	83.1%	0.661
Multinomial Naive Bayes	**85.2%**	**0.692**	**84.6%**	**0.663**
Support Vector Machines	83.2%	0.661	83.7%	0.654
C4.5 Decision Tree	82.6%	0.644	81.2%	0.576

Classification using user name and screen name

The results previously obtained with the *user name* and *screen name* features are described in detail in [55]. The 192 features allow to infer gender when the user self-assigns a name either in the *user name* or the *screen name*. With Multinomial Naive Bayes, the achieved accuracy was of 97.9% for English users and of 98.3% for Portuguese users. In our previous work, the purpose was to infer gender using only *screen name* and *user name*. For that reason, the data was biased and only users with a name in one of the *user name* and *screen name* fields were considered. 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.

Multinomial Naive Bayes achieved the best performance for both languages, 85.2% of accuracy for the English users and 84.6% for the Portuguese users. It is coherent with the results obtained in [55]. Though the Portuguese dataset has a higher baseline, the percentage of users with features is inferior to the English dataset. Results achieved for each of the methods are summarized in Table 4.

Classification using the user description

To evaluate the description features, we used the English dataset split in three subsets as previously described. The description field is not mandatory and from the 65k English users, only 79% have a description. This classifier only sends an output to the combined classifier if the user has a description. For the experiments, we consider all users, even the ones without description.

The used data was preprocessed as explained before. In order to test the classifiers, neither stopwords were removed nor stemming was performed. The representation of train, validation and test subsets was of ngrams with term frequency-inverse document frequency (TF-IDF) conversion and normalizing word frequencies. We applied dimensionality reduction, because the descriptions of all users are represented by thousands of tokens, making the classification task difficult. There are two approaches for dimensionality reduction:

1. **Feature reduction**, mapping the original list of attributes to a more compact representation. New attributes will combine original information sharing common statistical properties. Feature reduction can be obtained using methods like Singular Value Decomposition (SVD), Latent Semantic Analysis (LSA) or Principal Component Analysis (PCA)

2. **Feature selection**, selecting from the original list of attributes only a subset. Feature selection can be obtained using methods like Information Gain or Chi-square.

Table 5. Gender classification results for description features of English users.

	Accuracy	Kappa	Precision	Recall	F-Measure
Baseline	51.8%				
Logistic Regression	60.7%	0.200	63.0%	60.7%	0.580
Multinomial Naive Bayes	**61.6%**	0.225	61.7%	61.6%	0.611
Support Vector Machines	60.0%	0.182	63.8%	60.0%	0.566
C4.5 Decision Tree	58.9%	0.164	60.5%	58.9%	0.563

Being simpler and less time consuming, we used feature selection with the evaluator Information Gain and the search algorithm Ranker having the threshold property equal to zero.

A number of different parameters was tested and optimized, but the best performance was achieved using word unigrams, bigrams and trigrams combined, keeping 10000 instances. Table 5 shows the results obtained. Multinomial Naive Bayes achieved the best performance with an accuracy of 61.6%. The performance would be higher if only users with description were analyzed, but for our purpose, is necessary to analyze all users. These results are consistent with the work of [9], where the description is the less gender indicative field.

Some of the most strong description features of English users are similar to those presented by [9] or [56]. 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 semanticaly related with sports or interjections, as *man* or *bro*.

Classification using tweets content

For the experiments using tweets content, we will use the English dataset split in three subsets as previously described. The last 100 tweets from each user were extracted and the tweets text was preprocessed as explained previously.

Textual ngram features

To evaluate textual ngram features we used unigrams, bigrams, trigrams and the combination of the three. In order to test the classifiers, neither stopwords were removed nor was performed stemming. Different parameters were tested and optimized. Dimensionality reduction, TF-IDF conversion and normalizing word frequencies increased accuracy in the classifiers. We used feature selection with the evaluator Information Gain and the search algorithm Ranker having the threshold property equal to zero. 1000 ngrams were select for each algorithm. 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*.

Table 6 shows the results obtained using the previously described parameters. Column "Time (s)" contains the time spent to build each model. Support Vector Machine using unigrams achieves the highest performance, obtaining an accuracy of 73.8%. 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. We experimented

Table 6. Gender classification results for textual features of English users.

	Order	Time(s)	Accuracy	Kappa	Precision	Recall	F-measure
Baseline			51.8%				
C4.5	1	1165	60.1%	0.199	60.0%	60.0%	0.600
	2	1033	57.4%	0.146	57.4%	57.3%	0.574
	3	696	59.1%	0.186	59.7%	59.1%	0.589
	1-3	725	59.0%	0.177	58.9%	58.9%	0.589
LR	1	157	73.5%	0.468	73.5%	73.5%	0.734
	2	218	69.1%	0.380	69.1%	69.1%	0.691
	3	183	64.4%	0.287	64.4%	64.4%	0.644
	1-3	539	73.2%	0.463	73.2%	73.2%	0.732
MNB	1	119	**71.7%**	0.433	71.7%	71.7%	0.717
	2	166	68.6%	0.371	68.6%	68.6%	0.686
	3	150	62.4%	0.246	62.4%	62.4%	0.623
	1-3	244	71.6%	0.431	71.6%	71.6%	0.716
SVM	1	8824	**73.8%**	0.474	73.8%	73.8%	0.737
	2	2637	69.1%	0.382	69.1%	69.1%	0.691
	3	1910	64.3%	0.287	64.4%	64.3%	0.644
	1-3	13187	73.3%	0.464	73.3%	73.3%	0.732

Table 7. Gender classification results for textual ngram features of English users using Multinomial Naive Bayes.

Order	Tokens	Time (s)	Accuracy	Kappa
1	1000	26	71.7%	0.433
1	10000	30	72.8%	0.452
1	50000	40	71.3%	0.425
1	100000	34	71.2%	0.421
1-3	1000	213	**71.6%**	0.431
1-3	10000	236	73.0%	0.459
1-3	50000	224	73.1%	0.460
1-3	100000	259	**73.2%**	0.462

Multinomial Naive Bayes using the same parameters but without feature selection. Table 7 shows the results. Using a combinations of unigrams, bigrams and trigrams, the performance of Multinomial Naive Bayes constantly increased when more tokens were considered. A performance of 73.2% was achieved using 100k tokens. The time necessary to build a model, even when using 100k tokens is much inferior when comparing to Support Vector Machine algorithm. The time necessary to build a model depends on the availability of the processor and memory of the computer. We can observe the same Multinomial Naive Bayes experiences, took longer in our first experiments. Building a Multinomial Naive Bayes model with unigrams and 1000 tokens lasted 119 seconds in the first experiments, but only 26 seconds in the experiments where only Multinomial Naive Bayes was used.

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

Table 8. Labelled subsets of United Kingdom and United States users.

Subset	Users	Train	Test
United States	41034	31036	9998
United Kingdom	5780	4294	1486

Table 9. Gender classification results for textual ngram features of English users using geographical context.

	Subset	Time (s)	Accuracy	Kappa	Precision	Recall	F-Measure
Baseline			51.8%				
LR	All	539	73.2%	0.463	73.2%	73.2%	0.732
	UK	33	71.9%	0.421	71.8%	71.9%	0.717
	US	503	**73.8%**	0.471	73.7%	73.8%	0.737
MNB	All	9315	71.6%	0.431	71.6%	71.6%	0.716
	UK	174	72.7%	0.453	72.8%	72.7%	0.728
	US	248	**74.0%**	0.474	74.3%	74.0%	0.740
SVM	All	13187	73.3%	0.464	73.3%	73.3%	0.732
	UK	69	72.3%	0.429	72.1%	72.3%	0.721
	US	10997	**74.2%**	0.479	74.2%	74.2%	0.741

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%. We split the subsets in train and test as described in Table 8.

Due to the poor results obtained in the previous tests, we excluded the C4.5 decision tree algorithm. We used the same parameters from the experiences performed in the complete English dataset and used the combination of unigrams, bigrams and trigrams. Table 9 describes the results obtained. Creating models based on geography improved almost all algorithms accuracy. 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 algorithms. 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 algorithms.

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. Stopwords were not removed nor did we perform stemming. Dimensionality reduction, TF-IDF conversion and normalizing word frequencies were applied. We used feature selection with the evaluator Information Gain and the search algorithm Ranker having the threshold property equal to zero. 1000 tokens were select for each algorithm.

Table 10 shows the results of the textual ngram features in the Portuguese dataset. SVM and Multinomial Naive Bayes obtain an accuracy of about 93%. Logistic regression achieves 84.8% of accuracy. The accuracy achieved completely outperforms the results of 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. Again, the results obtained in the Portuguese dataset outperform the results from the English dataset.

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

Table 10. Gender classification results for textual ngram features of Portuguese users.

	Order	Accuracy	Kappa	Precision	Recall	F-Measure
Baseline		57.2%				
LR	1	84.2%	0.601	84.8%	84.2%	0.832
	3	76.5%	0.391	76.0%	76.5%	0.744
	1-3	**84.8%**	0.624	84.9%	84.8%	0.841
	1-4	82.1%	0.551	82.0%	82.1%	0.810
MNB	1	90.9%	0.789	90.9%	90.9%	0.909
	3	90.1%	0.762	90.3%	90.1%	0.899
	1-3	89.6%	0.771	90.7%	89.5%	0.898
	1-4	**93.3%**	0.847	93.3%	93.3%	0.933
SVM	1	88.2%	0.714	88.2%	88.2%	0.878
	3	81.7%	0.546	81.4%	81.7%	0.808
	1-3	89.6%	0.749	89.6%	89.5%	0.893
	1-4	**93.5%**	0.851	93.5%	93.5%	0.935

Table 11. Gender classification results using profile picture.

	Accuracy		
Dataset	Baseline	All data	Face detected
English	51.8%	67.2%	76.9%
Portuguese	57.2%	75.8%	85.7%

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.

Table 11 shows the results obtained using facial gender detection on both datasets. 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 baselines presented are from the complete dataset. The profile picture proved to be useful for gender detection.

Combined classifier

In the previous subsections, we evaluated the separate classifiers. A summary of the results obtained is shown in Fig. 8. In the English dataset, the *user name* and *screen name* features reach the highest accuracy 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 [9] 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.

In this section we will evaluate the accuracy of the combined classifier both with

Figure 8. Separate classifiers' accuracy results.

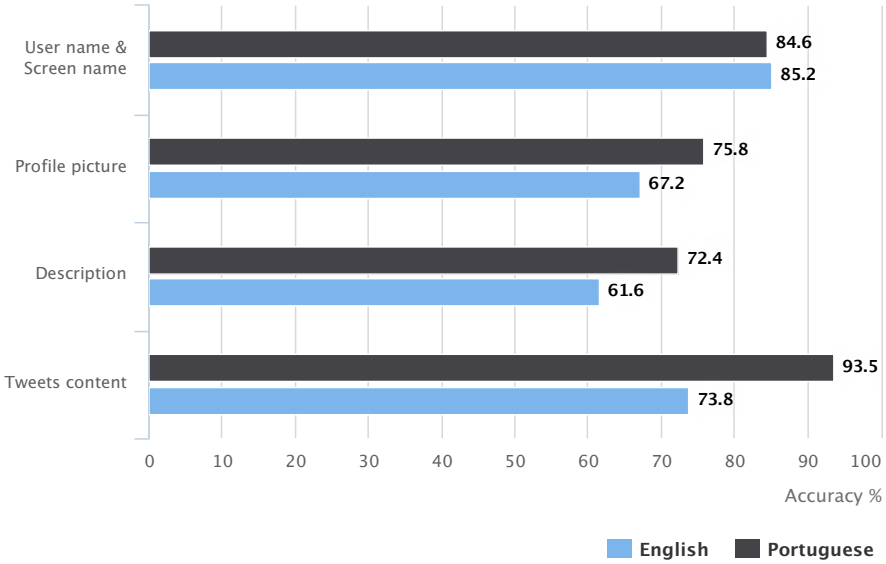


Table 12. Gender classification accuracy using the combined classifier.

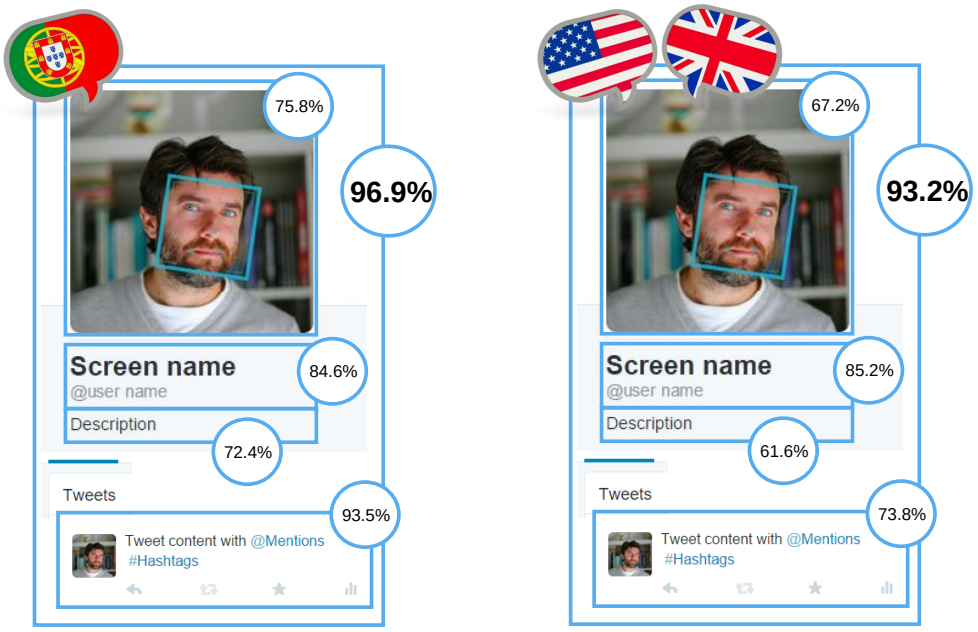
Dataset	Baseline (majority vote)	Combined
English	51.8%	93.2%
Portuguese	57.2%	**96.9%**

English and Portuguese users. The combined classifier 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 Support Vector Machine algorithm 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 following parameters: C=1.0 (complexity), epsilon=1.0E-12, kernel=PolyKernel.

Table 12 shows the accuracy obtained in both datasets using the combined classifier. The combined classifier improves the performance in 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. Fig. 9 summarizes the achieved accuracies per classifier for both datasets.

Figure 9. Classification accuracy per group of features for both datasets.



Conclusions

In this study, we experimented a method for gender detection using a combined classifier. Instead of applying the same classifier for all features, we grouped related features and classified them separately. The output of each feature was then used as input for the final combined classifier. We used extended labelled datasets from our previous works [45, 55], partitioned into train, validation and test subsets. The features, based on the users' content and profile information, were distributed in the following groups: user name and screen name, description, tweet content, profile picture and social network. The first group of features to be evaluated was *user name* and *screen name*. We used the 192 *user name* and *screen name* features from our first experiments. Multinomial Naive Bayes achieved the best performance for both languages, 85.2% of accuracy for the English users and 84.6% for the Portuguese users. For the classification using the user description features, the best performance was achieved using unigrams, bigrams and trigrams combined, keeping 10k instances. Again, Multinomial Naive Bayes achieved the best performance with an accuracy of 61.6%. For the classification using tweets' content, we extracted textual ngram features and style and sociolinguistic features. Support Vector Machine obtain an accuracy of about 73% for the English dataset and 93% for the Portuguese dataset. The performance of the English classifier improved to 74% when the experiments were made using only users from a specific region, in the case, the United States. The evaluation of the profile picture feature was done through the use of the Face++ API. The performance was 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 (not all users have a profile picture with a face). 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. Finally, the social network features were discarded, since no differences were observed when using these features. After the experiments of the separate classifiers, the predictions were retrieved and sent as inputs for the combined classifier. The prediction from the separate classifiers were only sent if

information was available. E.g.: if a user had no description, the input from that classifier would be empty. In the Portuguese dataset we obtained an accuracy of 96.9%. 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 the features proposed and using the combined classifier, one tweet could be 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.

We conclude by stating that we have reached our goal: we successfully built a combined classifier for Portuguese users and a classifier for English user, obtaining a high accuracy on both classifiers. Using our methodology, models can be built for other languages. To our best knowledge, we provide 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. 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. 2011;5:31. (document)

2. Conover M, Ratkiewicz J, Francisco M, Gonçalves B, Menczer F, Flammini A. Political polarization on twitter. In: ICWSM; 2011. . (document)

3. Culotta A. Detecting influenza outbreaks by analyzing Twitter messages. arXiv preprint arXiv:10074748. 2010;. (document)

4. Earle P, Guy M, Buckmaster R, Ostrum C, Horvath S, Vaughan A. OMG earthquake! Can Twitter improve earthquake response? Seismological Research Letters. 2010;81(2):246–251. (document)

5. Vieweg S, Hughes AL, 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. ACM; 2010. p. 1079–1088. (document)

6. Imran M, Castillo C, Diaz F, Vieweg S. Processing Social Media Messages in Mass Emergency: A Survey. ACM Computing Surveys (CSUR). 2015;47(4):67. (document)

7. Carvalho JP, Pedro V, Batista F. Towards Intelligent Mining of Public Social Networks' Influence in Society. In: IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS). Edmonton, Canada; 2013. p. 478 – 483. (document)

8. Leetaru K, Wang S, Cao G, Padmanabhan A, Shook E. Mapping the global Twitter heartbeat: The geography of Twitter. First Monday. 2013;18(5). (document)

9. Burger JD, Henderson J, Kim G, Zarrella G. Discriminating Gender on Twitter. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. EMNLP '11. Stroudsburg, PA, USA: Association for Computational Linguistics; 2011. p. 1301–1309. (document)

10. Koppel M, Argamon S, Shimoni AR. Automatically categorizing written texts by author gender. Literary and Linguistic Computing. 2002;17(4):401–412. (document)

11. Argamon S, Koppel M, Fine J, Shimoni AR. Gender, genre, and writing style in formal written texts. TEXT-THE HAGUE THEN AMSTERDAM THEN BERLIN-. 2003;23(3):321–346. (document)

12. Argamon S, Koppel M, Pennebaker JW, Schler J. Automatically profiling the author of an anonymous text. Communications of the ACM. 2009;52(2):119–123. (document)

13. Goswami S, Sarkar S, Rustagi M. Stylometric analysis of bloggers age and gender. In: Third International AAAI Conference on Weblogs and Social Media; 2009. . (document)

14. Koppel M, Schler J, Argamon S. Computational methods in authorship attribution. Journal of the American Society for information Science and Technology. 2009;60(1):9–26. (document)

15. Mukherjee A, Liu B. Improving gender classification of blog authors. In: Proceedings of the 2010 conference on Empirical Methods in natural Language Processing. Association for Computational Linguistics; 2010. p. 207–217. (document)

16. Cheng N, Chandramouli R, Subbalakshmi K. Author gender identification from text. Digital Investigation. 2011;8(1):78–88. (document)

17. Peersman C, Daelemans W, Van Vaerenbergh L. Predicting age and gender in online social networks. In: Proceedings of the 3rd international workshop on Search and mining user-generated contents. ACM; 2011. p. 37–44. (document)

18. Goswami S, Shishodia MS. A fuzzy based approach to stylometric analysis of blogger's age and gender. In: Hybrid Intelligent Systems (HIS), 2012 12th International Conference on; 2012. p. 47–51. (document)

19. Baumann A, Krasnova H, Veltri NF, Ye Y. Men, Women, Microblogging: Where Do We Stand? Proceedings der 12 Internationalen Tagung Wirtschaftsinformatik. 2015;. (document)

20. Holmes J, Meyerhoff M. The handbook of language and gender. vol. 25. John Wiley & Sons; 2008. (document)

21. Eckert P, McConnell-Ginet S. Language and gender. Cambridge University Press; 2013. (document)

22. Bucholtz M, Hall K. Identity and interaction: A sociocultural linguistic approach. Discourse studies. 2005;7(4-5):585–614. (document)

23. Fischer JL. Social influences on the choice of a linguistic variant. Word. 1958;14(1):47–56. (document)

24. Labov W. The social stratification of English in New York city. Cambridge University Press; 2006. (document)

25. Schler J, Koppel M, Argamon S, Pennebaker JW. Effects of Age and Gender on Blogging. In: AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. vol. 6; 2006. p. 199–205. (document)

26. Aravantinou C, Simaki V, Mporas I, Megalooikonomou V. Gender Classification of Web Authors Using Feature Selection and Language Models. In: Ronzhin A, Potapova R, Fakotakis N, editors. Speech and Computer. vol. 9319 of Lecture Notes in Computer Science. Springer International Publishing; 2015. p. 226–233. (document)

27. 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. SMUC '10. New York, NY, USA: ACM; 2010. p. 37–44. (document)

28. Al Zamal F, Liu W, Ruths D. Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors. ICWSM. 2012;270. (document)

29. McPherson M, Smith-Lovin L, Cook JM. Birds of a feather: Homophily in social networks. Annual review of sociology. 2001;p. 415–444. (document)

30. Mislove A, Lehmann S, Ahn YY, Onnela JP, Rosenquist JN. Understanding the Demographics of Twitter Users. ICWSM. 2011;11:5th. (document)

31. 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. . (document)

32. Bamman D, Eisenstein J, Schnoebelen T. Gender in Twitter: Styles, stances, and social networks. CoRR abs/12104567. 2012;. (document)

33. Deitrick W, Miller Z, Valyou B, Dickinson B, Munson T, Hu W. Gender identification on Twitter using the modified balanced winnow. Communications and Network. 2012;4(3). (document)

34. Miller Z, Dickinson B, Hu W. Gender prediction on twitter using stream algorithms with N-gram character features. International Journal of Intelligence Science. 2012;2(4A). (document)

35. Fink C, Kopecky J, Morawski M. Inferring Gender from the Content of Tweets: A Region Specific Example. In: ICWSM; 2012. . (document)

36. Joachims T. Making large scale SVM learning practical. Universität Dortmund; 1999. (document)

37. Pennebaker JW, Chung CK, Ireland M, Gonzales A, Booth RJ. The development and psychometric properties of LIWC2007; 2007. (document)

38. 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. . (document)

39. 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; 2013. p. 1010–1019. (document)

40. Ciot M, Sonderegger M, Ruths D. Gender Inference of Twitter Users in Non-English Contexts. In: EMNLP; 2013. p. 1136–1145. (document)

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

42. Ludu PS. Inferring gender of a Twitter user using celebrities it follows. arXiv preprint arXiv:14056667. 2014;. (document)

43. Merler M, Cao L, Smith JR. 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. IEEE; 2015. p. 1–6. (document)

44. Vicente M. Detecting Portuguese and English Twitter users' gender. ISCTE-IUL - Instituto Universitário de Lisboa; 2015. (document)

45. Vicente M, Batista F, Carvalho JP. Twitter gender classification using user unstructured information. In: Proc. of IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). Istambul, Turkey; 2015. . (document)

46. Brogueira G, Batista F, Carvalho JP, Moniz H. Expanding a Database of Portuguese Tweets. In: Pereira MJV, Leal JP, Simoes A, editors. 3rd Symposium on Languages, Applications and Technologies. vol. 38 of OpenAccess Series in Informatics (OASIcs). Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik; 2014. p. 275–282. (document)

47. Bechar-Israeli H. FROM< Bonehead> TO< cLoNehEAd>: NICKNAMES, PLAY, AND IDENTITY ON INTERNET RELAY CHAT1. Journal of Computer-Mediated Communication. 1995;1(2):0–0. (document)

48. Calvert SL, Mahler BA, Zehnder SM, Jenkins A, Lee MS. Gender differences in preadolescent children's online interactions: Symbolic modes of self-presentation and self-expression. Journal of Applied Developmental Psychology. 2003;24(6):627–644. (document)

49. Stopczynski A, Sekara V, Sapiezynski P, Cuttone A, Madsen MM, Larsen JE, et al. Measuring large-scale social networks with high resolution. PloS one. 2014;9(4):e95978. (document)

50. Baptista J, Batista F, Mamede NJ, Mota C. Npro: um novo recurso para o processamento computacional do Português. In: XXI Encontro APL; 2005. . (document)

51. Corney MW. Analysing e-mail text authorship for forensic purposes. Queensland University of Technology; 2003. (document)

52. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA Data Mining Software: An Update. SIGKDD Explor Newsl. 2009 Nov;11(1):10–18. (document)

53. Feldman R, Sanger J. The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge University Press; 2007. (document)

54. Harris ZS. Distributional structure. Word. 1954;. (document)

55. Vicente M, Carvalho JP, Batista F. Using Unstructured Profile Information for Gender Classification of Portuguese and English Twitter users. In: Proc. of Symposium on Languages, Applications and Tecnologies (SLATE'15). short papers. Madrid, Spain; 2015. . (document)

56. Van Zegbroeck E. Predicting the Gender of Flemish Twitter Users Using an Ensemble of Classifiers. 2014;. (document)