Department of Information Science and Technology

Detecting Portuguese and English Twitter users' gender

Marco Paulo Fernandes Vicente

A dissertation submitted in partial fulfillment of the requirements for the degree of **Master in Open Source Software**

Supervisor:
PhD Fernando Batista, Assistant Professor,
ISCTE-IUL – Instituto Universitário de Lisboa

Co-Supervisor:
PhD João Paulo Carvalho, Assistant Professor,
Instituto Superior Técnico – Universidade de Lisboa

October 2015

Resumo

Os serviços de redes sociais existentes proporcionam meios para as pessoas comunicarem e exprimirem os seus sentimentos de uma forma fácil. O conteúdo gerado por estes utilizadores contém indícios dos seus comportamentos e preferências, bem como outros metadados que estão agora disponíveis para investigação científica. O Twitter em particular, tornou-se uma fonte importante para estudos das redes socias, sobretudo porque fornece um modo simples para os utilizadores expressarem os seus sentimentos, ideias e opiniões; disponibiliza o conteúdo gerado pelos utilizadores e os metadados associados à comunidade; e fornece interfaces web e interfaces de programação de aplicações (API) para acesso aos dados de fácil utilização. Para muitos estudos, a informação disponível sobre um utilizador é relevante. No entanto, o atributo de género não é fornecido ao criar uma conta no Twitter.

O foco principal deste estudo é inferir o género dos utilizadores através da informação disponível. Propomos uma metodologia para a detecção de género de utilizadores do Twitter, usando informação não estruturada encontrada no perfil do Twitter, no conteúdo gerado pelo utilizador, e mais tarde usando a imagem de perfil do utilizador. Em estudos anteriores, um dos desafios apresentados foi a tarefa de etiquetar manualmente de dados, que revelou exigir bastante trabalho. Neste estudo, propomos um método para a criação de conjuntos de dados etiquetados de uma forma semi-automática, utilizando um conjunto de atributos com base na informação não estruturada de perfil. Utilizando os conjuntos de dados etiquetados, associamos conteúdo textual ao seu género e criamos modelos, com base no conteúdo gerado pelos utilizadores, e pela informação de perfil. Exploramos classificadores supervisionados e não supervisionados e avaliar os resultados em ambos os conjuntos de dados de utilizadores Portugueses e Ingleses do Twitter. Obtivemos uma precisão de 93,2% com utilizadores ingleses e uma precisão de 96,9% com utilizadores Portugueses. A metodologia proposta é independente do idioma, mas o foco foi dado a utilizadores Portugueses e Ingleses.

Abstract

Existing social networking services provide means for people to communicate and express their feelings in an easy way. Such user generated content contains clues of user's behaviors and preferences, as well as other metadata information that is now available for scientific research. Twitter, in particular, has become a relevant source for social networking studies, mainly because: it provides a simple way for users to express their feelings, ideas, and opinions; makes the user generated content and associated metadata available to the community; and furthermore provides easy-to-use web interfaces and application programming interfaces (API) to access data. For many studies, the available information about a user is relevant. However, the gender attribute is not provided when creating a Twitter account.

The main focus of this study is to infer the users' gender from other available information. We propose a methodology for gender detection of Twitter users, using unstructured information found on Twitter profile, user generated content, and later using the user's profile picture. In previous studies, one of the challenges presented was the labor-intensive task of manually labelling datasets. In this study, we propose a method for creating extended labelled datasets in a semi-automatic fashion. With the extended labelled datasets, we associate the users' textual content with their gender and created gender models, based on the users' generated content and profile information. We explore supervised and unsupervised classifiers and evaluate the results in both Portuguese and English Twitter user datasets. We obtained an accuracy of 93.2% with English users and an accuracy of 96.9% with Portuguese users. The proposed methodology of our research is language independent, but our focus was given to Portuguese and English users.

Palavras Chave
Keywords

Palavras chave

Mineração de Texto

Classificação de Género

Utilizador Twitter

Selecção de Atributos

Classificação de Texto

Keywords

Text Mining

Gender Classification

Twitter User

Feature Selection

Text Classification

Acknowledgements

MOSS @MOSS
Thanks to all professors and colleagues for the great intellectual camaraderie and sharing of ideas #MOSS #OpenSource

Gaspar Brogueira @GasparBrogueira
Always available and seeking to push the boundaries. Helped both with work and ideas. #MOSS #OpenSource #Twitter #BigData #API

Carlos J. Costa @carlosjcosta
MOSS saturday meetups were really useful #MOSS #OpenSource #ACMStudentChapter

Manuela Aparicio @ManuelaAparicio
Spent personal time motivating and guiding us #MOSS #Thesis #RelatedWork

João Paulo Carvalho @ukedone
Co-advisor with great insights and pragmatism that greatly improved the quality of the produced papers and this dissertation #Fuzzy #Madrid #Istanbul

Fernando Batista @ferbatis
Thanks for the teaching, dedication and constant support. For helping in both the small and the big decisions #WEKA #Python #NLP

Helena Vicente @hfvicente
Super sister!
👩🏽👩🏽👩🏽👩🏽👩🏽
#Twitter #Users #Labelling

Vicente @Vicentes
[1/3] To my family for the support and the understanding.
Specially my beloved mother, my sisters and my missed father

M. Vicente @BabyM
[2/3] To my daughter, my main motivation and joy. "Vai trabalhar, pai..."

Paulos Vicente @familyOfThree
[3/3] Thanks to my wonderful wife, for her support, love, and for being responsible for the best years of my life.

Contents

List of Figures

List of Tables

Abbreviations

API	Application Programming Interface
ARFF	Attribute-Relation File Format
FCM	Fuzzy c-Means
IR	Information Retrieval
MNB	Multinomial Naive Bayes
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
SVM	Support Vector Machine
TF-IDF	Term Frequency – Inverse Document Frequency
URL	Uniform Resource Locator
WEKA	Waikato Environment for Knowledge Analysis

1 Introduction

Begin at the beginning, the King said, very gravely, and go on till you come to the end: then stop.

Lewis Carroll, *Alice in Wonderland*

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[1]. 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. Astronaut Mike Massimino sent the first tweet from space on 12 May 2009, as shown on Figure 1.1. Twitter has been influential in social events, like the Arab Spring (Lotan et al., 2011).

[1] Twitter usage [viewed 13 October 2015]. Available from: https://about.twitter.com/company

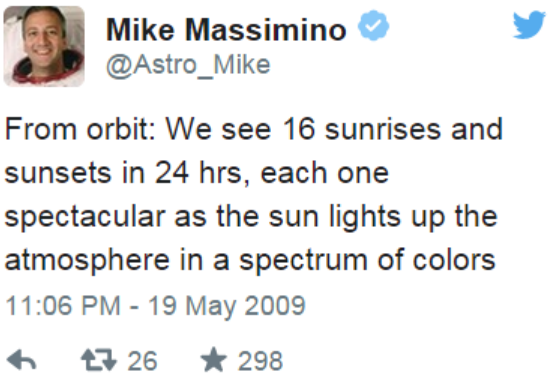


Figure 1.1: The first tweet from space, sent by astronaut Mike Massimino.

Being an enormous source of user-generated data, Twitter has become a major tool for social networking studies (Brogueira et al., 2014; Rosa et al., 2014). 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 (Conover et al., 2011) are now being extensively applied. Twitter is also being used to practical applications, such as to monitor diseases, e.g. detect flu outbreaks (Culotta, 2010), to improve response to natural catastrophes, e.g. detect earthquakes (Earle et al., 2010), or even to enhance awareness in emergency situations (Vieweg et al., 2010; Imran et al., 2015).

1.1 Definition of the problem and objectives

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 (Carvalho et al., 2013). 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 has a birthday field in the profile and tweets are georeferenced, 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. Female users are less likely to be

interested in that campaign. The gender information allows advertising to be effective and social studies to be more accurate.

The main objective of the proposed research is to automatically detect the gender of a Twitter user (male or female), based on the available information extracted from both the user's profile and tweets' content. The study involve both Portuguese and English users. A study of 46M georeferenced tweets, performed by Leetaru et al. (2013), reveals that Portuguese is the third most used language in Twitter with 6% of the georeferenced tweets. English is the most used language with 38% of the georeferenced tweets. Previous studies of gender classification have not focused in the Portuguese language. Burger et al. (2011) studied a multilingual corpora, but the results presented are global, and the accuracy for each language is not revealed.

The study has the following objectives:

- Investigate the state-of-the-art of Twitter gender detection;
- Create a method to semi-automatically label Twitter users' gender;
- Evaluate and improve the performance of the semi-automatic method;
- Propose a methodology to detect Twitter users gender;
- Assess the results of the proposed methodology.

A key objective of this study is to propose a methodology applicable to gender detection in other Indo-European languages.

1.2 Proposed methodology

The methodology used was based in the following set of steps:

- Research planning;
- Literature analysis and understanding of the state-of-the-art of the subject;
- Data extraction, preprocessing and labelling;
- Design, develop and test a methodology to gender detection;
- Document results and findings;
- Compare, review and discuss results.

1.3 Scientific contribution

This study presents the following contributions:

- a review of the existing approaches to overcome the problem of Twitter gender classification;
- proposes a novel methodology for the semi-automatic creation of gender labelled datasets;
- proposes a gender detection methodology based both on Twitter profile and tweets' content using a combined classifier;
- demonstrates the successful application of the methodology both in Portuguese and English users;
- introduces the first study of gender detection applied to Portuguese users.

The work conducted in this study has resulted in the following publications:

- Marco Vicente, Fernando Batista and Joao Paulo Carvalho (2015) Twitter gender classification using user unstructured information. In Proc. of IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). Istanbul, Turkey, Aug.
- Marco Vicente, Joao P. Carvalho and Fernando Batista (2015) Using Unstructured Profile Information for Gender Classification of Portuguese and English Twitter users. In Proc. of Symposium on Languages, Applications and Technologies (SLATE'15). Madrid, Spain, June.
- Marco Vicente, Joao P. Carvalho and Fernando Batista (2015) Using Unstructured Profile Information for Gender Classification of Portuguese and English Twitter users. In Languages, Applications and Technologies. 4th International Symposium, Slate 15, Madrid, Spain, June 18-19, 2015, Revised Selected papers. Communications in Computer and Information Science, vol. number 563 (to appear).

The first publication is related with the work described in Chapter 4, where supervised and unsupervised methods were used to gender detection, the second describes our efforts to create an approach for gender classification of both Portuguese and English Twitter users. The third publication corresponds to a post-proceedings publication of selected papers, where our second publication was selected and extended.

1.4 Structure of the dissertation

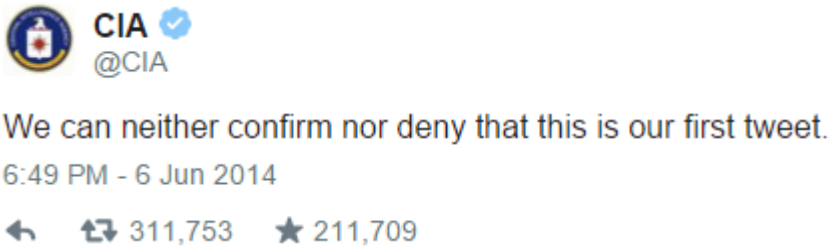
This dissertation presents the research undertaken as part of this thesis and suggests possible directions for future research. The document is structured as follows:

Chapter 2 makes an overview of social networks, particularly of Twitter. Chapter 3 provides an in-depth review of the current state of the art on Twitter gender detection, features currently in use, limitations of these works and describes our proposed approach.

Chapters 4, 5 and 6 follow a chronological perspective. Chapter 4 reports on our initial work concerning gender detection using only unstructured profile information. Chapter 5 discusses the creation of semi-automatic labelled datasets, using the features proposed in Chapter 4 and extending the datasets. Chapter 6 presents our methodology for gender detection, reports different experiments comparing different supervised and unsupervised methods and discusses the obtained results. Finally, Chapter 7 presents the conclusions and proposes a number of future tasks to further extend the work here described.

Background

2



This chapter describes some relevant concepts for our study. In Section 2.1, we describe social networks and microblogs. Next, in Section 2.2, we describe Twitter and associated concepts.

2.1 Social networks

A social network is a social construction made of nodes, usually composed by individuals or organizations (Newman, 2003; Hanneman and Riddle, 2005). It specifies the ways in which people and organizations are connected socially, ranging from casual connection to familiar bounds (Jamali and Abolhassani, 2006). Social networking allows to increase personal and business connections through online communities. Creating interconnected online communities, social networking allows individuals and organizations to make contacts that would be improbable otherwise. Depending on the social media platform, users may contact any other users. In some cases, users can only contact anyone they have a connection to, and consequently anyone that contact has a connection to, and so on. Some services require users to have a previous connection to contact other users. Social networks like LinkedIn are called vertical social networks, as they bring together individuals who share a particular subject or interest. In the case of LinkedIn, the shared interest is professional, allowing companies and professionals to connect with each other.

Microblog

Microblogging is a particular kind of social networking service. A form of blog publishing that allows users to share short text updates, usually less than 200 characters, and post them to be seen publicly or by only a selected group chosen by the user. These texts are sent by a variety of means such as short messaging services (SMS), instant messaging, email or web. The use of a blog is considered "micro" when it allows the insertion of text up to 200 characters or less. One of the most popular microblogging sites today is Twitter. The free service allows users to post data through SMS, email or by the web. Initially, the main focus of the microblogging messages was to share microposts related to daily activities of users. The initial slogan of Twitter "What are you doing?", the most popular service in the category, is based on the daily life activities. However, individuals and companies use microblogging as a way to showcase their blogs or businesses. A new trend in Twitter is the use of microblogging as a new journalism format (Bei, 2013). The advantages of microblogs are how easy they are to use, the brevity of the texts, the user mobility and the virtual networks they allow to create, especially for those who want increase the visibility of their traditional online presence.

2.2 Twitter

Twitter is a microblogging service that allows users to send updates and read updates from other people or organizations from the user's contacts. Status updates, known as "tweets", consist of text up to 140 characters and can be sent via Web, SMS or Mobile Application. The updates are displayed in the user profile in real time and also sent to other users who have signed to receive them. Users can receive updates from a profile through the official website, Rich Site Summary (RSS), Short Message Service (SMS) or specialized software. Twitter was created in 2006, and soon gained extensive notability and popularity worldwide. According to Statista[1], Portuguese is the fifth most used language on Twitter and English is the most used language. Twitter has several means to share information, like retweets, where a user replicates a particular message from other user to the list of followers, giving credit to the original author, trending topics, or TTs, which are a real-time list of most posted words in Twitter around the world, and can be divided by countries. It is a real-time contact tool. Twitter has become well known because celebrities use microblogging to communicate with their fans. Figure 2.1 shows a selfie shared at the 2014 Oscars Ceremony: The tweet was retweeted over 2 million times less than 24 hours later.

[1]Most-used languages on Twitter as of September 2013 [viewed 13 October 2015]. Available from: http://www.statista.com/statistics/267129/most-used-languages-on-twitter/

Figure 2.1: Selfie on the Oscars - the most retweeted image on Twitter.

Twitter may or may not be synchronous, all messages are recorded on each user timeline and they answer when they want to. The dialogues are public. Anyone can monitor user's conversations. A user can view the contacts of its contacts.

Twitter does not request or store much information about their users. A Twitter profile, as illustrated in Figure 2.2, only requires a *user name* (e.g. @POTUS) and an associated *email*. The remaining profile information is optional, namely:

- Header photo
- Profile photo
- Screen Name. E.g.: President Obama
- Bio (maximum 160 characters).
- Location. E.g.: Washington DC
- Website
- Theme color
- Birthday (available since 6 July 2015[2])

[2]Celebrate your birthday on Twitter [viewed 13 October 2015]. Available from: https://blog.twitter.com/2015/hbd-celebrate-your-birthday-on-twitter

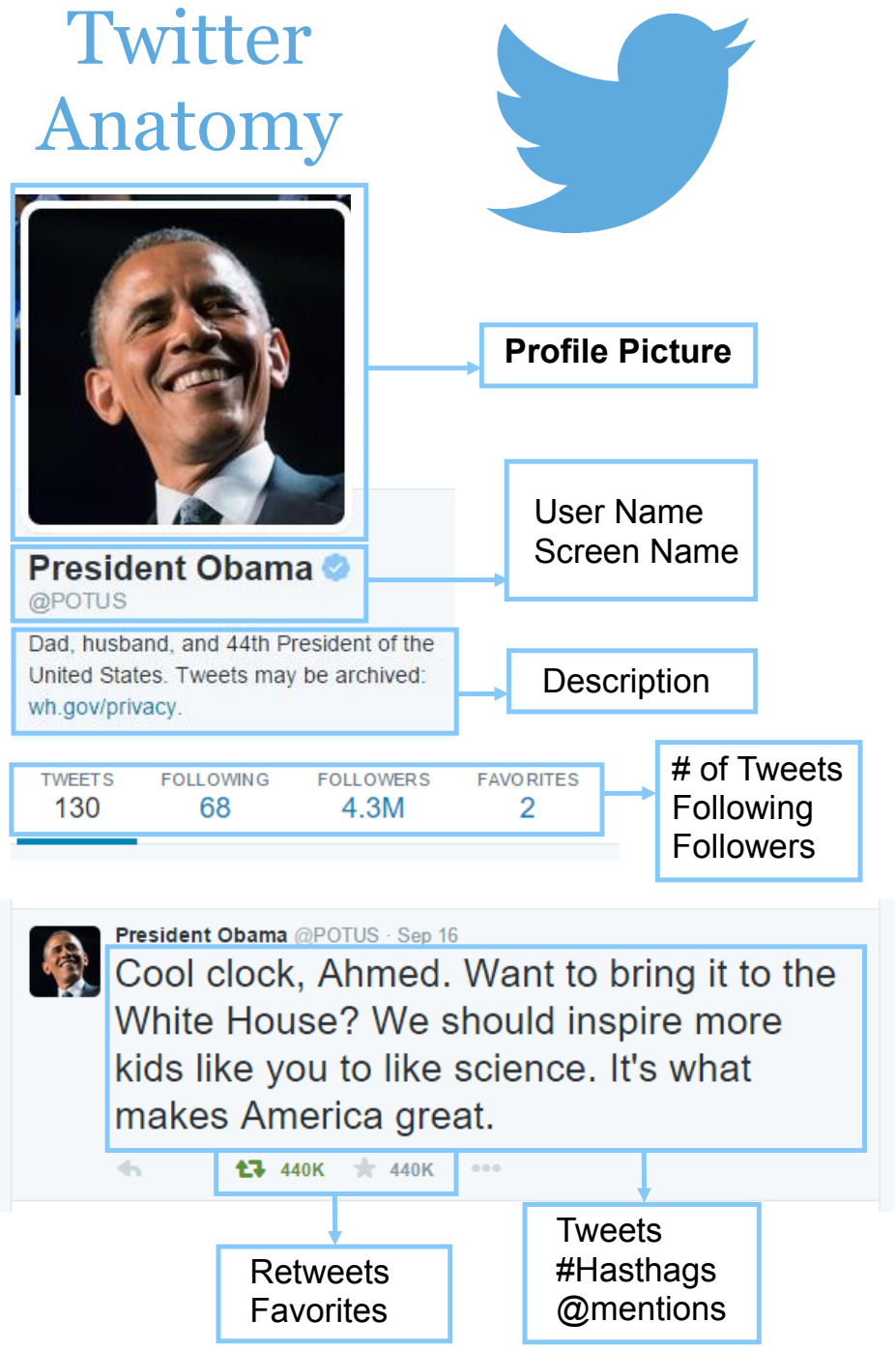


Figure 2.2: Twitter Anatomy.

To better understand Twitter, it is necessary to know some concepts associated with the microblogging service. Here, we present a list of terms used in Twitter context:

Tweeple/tweeps - Twitter users.

Tweet - Short 140-character message or status update sent on Twitter. Tweet content can have pictures, URLs, hashtags and mentions. Tweets can have up to 140 characters, including shared URLs.

Retweet (RT) - Retweet, abbreviated RT, is a status update reposted by a user.

Timeline - The stream of tweets of users and organizations followed by the user. The tweets are presented in a chronological order.

Following - Tweets are public. Users choose whose tweets want to receive by following other users. Unlike other social networking services, following does not have to be mutual. Users can follow other users without being followed back.

Followers - Users who follow a particular user and see this user's tweets and retweets in his timeline.

Hashtag (#) - A hashtag marks a keyword or topic and is prefixed by the symbol # (e.g.: #ObamaCare). Twitter search uses hashtags to categorize tweets.

Mention (@) - A mention is a sign "@" followed by a user name in a tweet or retweet and is used for mentioning or replying to other users.

URL Shortening - Twitter automatically uses its URL shortener, since the URL counts for the limit of 140 characters.

Geolocation - Users can add geolocation information to tweets. This information is useful to share where the user was when posted the tweet.

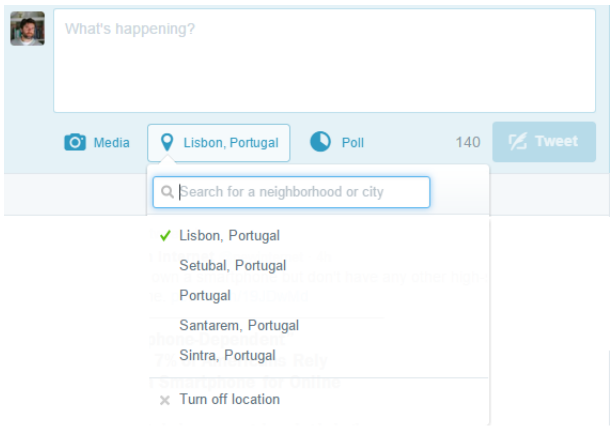


Figure 2.3: Geolocating a tweet.

Direct Message (DM) - Users can send private (direct) messages to other users. In order to send a direct message, the recipient has to be following the user sending the direct message.

Due to the 140-characters limitation, Twitter language usage differs from regular social networks, as Facebook, or LinkedIn. This limitation forces users to post short messages, but also provides a fast way to communicate. A Twitter user can post up to 1000 tweets per day.

Related work

3

The thing that hath been, it is that which shall be; and that which is done is that which shall be done: and there is no new thing under the sun.

Ecclesiastes 1:9

In this chapter, we review the various approaches that have been proposed for the problem of gender detection on Twitter. In Section 3.1 we start with an overview of the problem of gender detection. In Section 3.2 we review the existing studies for the problem of gender detection on Twitter. Finally, in Section 3.3, we discuss the main challenges faced by researchers detecting the gender of Twitter users, what solutions have been proposed, and how our approach both resembles and differs from the existing studies.

3.1 Gender classification

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 (Koppel et al., 2002; Argamon et al., 2003, 2009; Goswami et al., 2009; Koppel et al., 2009; Mukherjee and Liu, 2010; Cheng et al., 2011; Peersman et al., 2011; Goswami and Shishodia, 2012; Baumann et al., 2015).

The study of the relation between gender and language usage is extensive (for an overview, see e.g.: Holmes and Meyerhoff (2008); Eckert and McConnell-Ginet (2013)). 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 (Bucholtz and Hall, 2005;

Fischer, 1958; Labov, 2006). Koppel et al. (2002), 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 (Schler et al., 2006), 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 Koppel et al. (2009) 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 Goswami et al. (2009). They improved the overall accuracy to 89.2%, using average sentence length, usage of slang and usage of non-dictionary words.

Cheng et al. (2011) studied gender identification using two large text datasets: 1) Reuters Corpus[1] Volume 1 newsgroup data, a large collection of Reuters News stories, and 2) Enron email dataset, containing emails from about 150 users, mostly senior management of Enron[2]. They applied three different supervised classification techniques (support vector machine, Bayesian logistic regression and AdaBoost decision tree), using linguistic and stylometric features, obtaining an accuracy of 85.1% on gender prediction using SVM.

Peersman et al. (2011) used a corpus of about 1.5M Flemish Dutch Netlog[3] posts for gender classification. 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.

Aravantinou et al. (2015) 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%. For an overview of the existing research analyzing the differences between gender in the usage of microblogs, see Baumann et al. (2015).

[1] http://trec.nist.gov/data/reuters/reuters.html
[2] See more here: https://www.cs.cmu.edu/~./enron/
[3] Netlog is a Belgian online social networking platform: http://nl.netlog.com/

3.2 Gender classification of Twitter users

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 Rao et al. (2010). Their goal was to infer latent user attributes, namely: gender, age, regional origin and political orientation. They manually annotated 500 users of each gender, crawling sources of sororities, fraternities, and male and female hygiene products.

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. The latter feature results contrast with the findings of Garera and Yarowsky (2009), studying conversation speeches using the Fisher telephone conversation corpus (Cieri et al., 2004), where analogous communication behavior features were highly gender distinctive.

Rao et al. (2010) defined the classification task as a binary classification problem. They built three supervised classification models. The first model, applied to their sociolinguistic features, was an SVM based binary classifier using the SVMLite package[4]. The second model, applied to unigrams and bigrams of tweets, was another SVM based binary classifier. The text was previously normalized and lowercased, preserving both emoticons and punctuation sequences. Finally, the third model was an SVM classifier stacking, using as features the predictions from the first two models.

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 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.

[4]http://svmlight.joachims.org/

The state-of-the-art study of Burger et al. (2011) collected a large multilingual dataset labelled with gender. While Rao et al. (2010) manually annotated 1000 English users, Burger et al. (2011) created a corpus of approximately 213M tweets from 18.5M Twitter users labelled with gender. This was not the first Twitter data corpora (Petrović et al., 2010; Eisenstein et al., 2010), but was the first with gender labelling. For the creation of the corpora, they extracted sampling data from Twitter API since 2009 till 2011. Next, they followed users who had filled the URL field in the profile information with blogging websites and sampled the corresponding profiles. They attributed the gender found on the corresponding website to the Twitter user. To validate the correctness of the method, they manually validated a sample of randomly selected 1000 Twitter users, by examining Twitter profile description. Only 15% of the sample had explicit gender information and in all the gender information agreed with the corresponding blog profile. The labelled users were divided in three datasets, training (147k users), development (18k users) and test (18k users). The final corpus was divided in 55% female and 45% male users. The most representative languages in the corpus are English (67%), Portuguese (14%) and Spanish (6%). This distribution is different from the distribution reported by Wauters (2010)[5], where English represented 50% of the tweets, Japanese 14%, Portuguese 9%, Malay 6% and Spanish 4%. Twitter users who also have a blogging website may have a different distribution from Twitter users as a whole. While not representing the Twitter users as a whole, sampling users having blogging websites may have filtered spam and bot accounts. English has decreased its preponderance on Twitter. Figure 3.1[6] illustrates the most used languages on Twitter, as of September 2013.

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.

The experiments were performed using SVMs, Naive Bayes and Balanced Winnow2 machine learning algorithms to build gender classification models. The study suggests a set of learning parameters to improve Winnow's performance. A usage of low learning rate (0.03) when analyzing only one attribute (e.g.: only description, or only tweet text features) and a higher learning rate (0.20) when using combined features. Also, the use of large margin (35%) and after each iteration reduce de learning rate by 50%. Balanced Winnow2 proved to be more accurate and less time consuming. 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.

[5]http://semiocast.com/downloads/Semiocast_Half_of_messages_on_Twitter_are_not_in_English_20100224.pdf
[6]Data: Semiocast (languages and tweets by country, from a 10 percent sample).

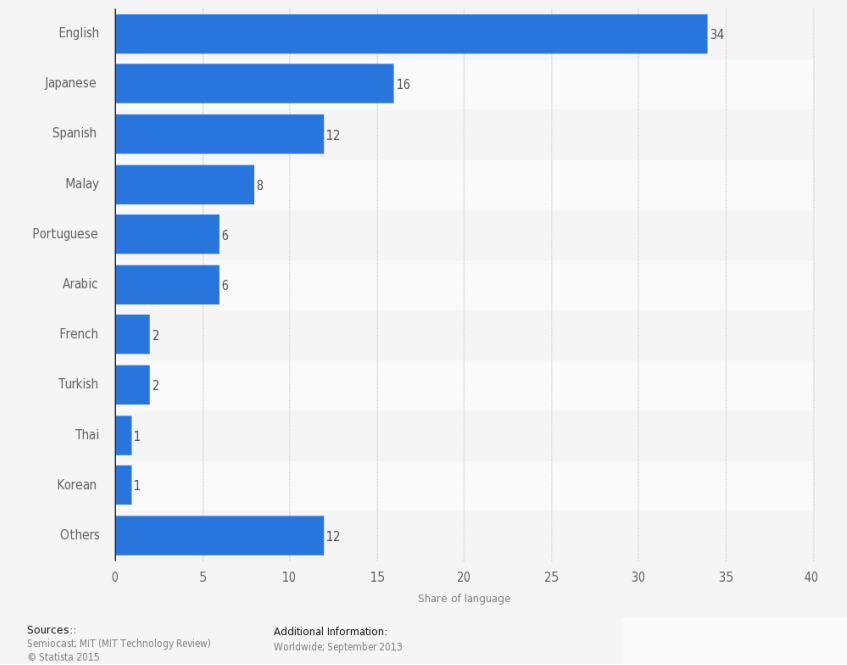


Figure 3.1: Most used languages on Twitter as of September 2013.

They further compared the automatic classification with a manual human task classification, using the Amazon Mechanical Turk (AMT). This was not the first effort to apply AMT to a set of natural language processing tasks (Callison-Burch and Dredze, 2010; Kaisser and Lowe, 2008). Five workers were assigned for each task, using both a majority vote approach and an expectation maximization algorithm. 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 Rao et al. (2010): 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, Al Zamal et al. (2012) 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 (McPherson et al., 2001).

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 Mislove et al. (2011)). 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 SVM-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 (Liu et al., 2012), 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.

Bamman et al. (2012) 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 Al Zamal et al. (2012), 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.

Deitrick et al. (2012) propose the use of neural network models for gender identification. Their limited dataset was composed of 3031 manually labelled tweets. They applied both Bal-

anced 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, Miller et al. (2012) 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.

Fink et al. (2012) present a region-specific study, focusing on Nigerian Twittter users. They label users based on the tweets' geolocation to create their dataset. Their experiments use SVM with a linear kernel implementation (Joachims, 1999), based on word unigrams, hashtags and Linguistic Inquiry and Word Count, or LIWC (Pennebaker et al., 2007). They report an accuracy of 81% when using unigrams as features.

While the previous studies focused on tweets' content alone, Liu and Ruths (2013) study the connection between gender and the self-reported first name. They add name features to tweets ngram features. Using an SVM classifier, they improved the accuracy from a baseline of 83%, using only ngrams, to 87%, using also first name features. Bergsma et al. (2013) studies gender detection based on users' preferences and location. They classify using distributed k-Means clustering and SVM 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 Burger et al. (2011) was multilingual, the classification was global and no data was given regarding the classification of separate languages. Ciot et al. (2013) 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"[7] 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. Halteren and Speerstra (2014) studied a corpus of Dutch tweets of 600 labelled users. Using tweet text only, using both character and token ngrams, they achieved an accuracy of 95.5% detecting gender. The machine learning system used was Sup-

[7]The French words "je suis" can be translated to "i am" in English

port Vector Regression with a 5-fold cross-validation on the corpus. In their research Linguistic Profiling and TiMBL are used, but with inferior accuracy.

Bamman et al. (2014) 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.

Ludu (2014) studies gender classification using celebrities the user follows as features combined with tweets content features. The accuracy achieved with SVM-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%.

Merler et al. (2015) 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%.

Previously mentioned studies improved the state-of-the-art, either suggesting new features or improving the performance of the existing classifiers. However, many other studies have focused on gender detection Alowibdi et al. (2013), Kokkos and Tzouramanis (2014), Ugheoke (2014), Nguyen et al. (2014), Alowibdi et al. (2014), Van Zegbroeck (2014), You et al. (2014), Jaech and Ostendorf (2015), Arroju et al. (2015).

3.3 Proposed approach

According to our previous analysis, gender detection on Twitter presents several challenges:

i) The inexistence of labelled corpora has imposed a labor intensive task to previous researchers. Consequently, some studies use small labelled datasets for the creation of their models. Usually the task of labelling is performed by looking for names in the profile information, either manually or using services like Amazon Mechanical Turk.

ii) Most of the previous work has been done for the English language. To our best knowledge, there is no study of gender detection for Portuguese users. Moreover, Ciot et al. (2013) showed that applying previous research in different languages might not result in similar results, suggesting the use of language-specific features in order to improve classifiers' accuracy.

iii) Previous research focus in textual features, seldom on pictures posted, but to the best of our knowledge an integrated use of features, using the profile picture, textual content and profile

information, has not yet been explored.

In this study, we propose a language-independent gender detection methodology based on Twitter profile information and on the content of tweets. We create gender labelled datasets that can be used in future research. We also propose the use of a combined classifier that works as follows: we create separate classifiers for each group of features and send the prediction of each classifier as input to our combined classifier. We apply our proposed methodology to both Portuguese and English users, though it is applicable to gender detection in other Indo-European languages.

Gender classification from profile information

4

What's in a name? that which we call a rose
By any other name would smell as sweet.

William Shakespeare, *Romeo and Juliet (II, ii, 1-2)*

This Chapter proposes a method to automatically detect the user's gender (male or female), uniquely based on unstructured information extracted from the user's profile, and made available by Twitter for each tweet. The only restriction for this method is that within the user profile there is at least a sequence of characters matching a name contained within a dictionary. A set of manually defined features is proposed for extracting useful information from the user's profile attributes, namely *user name,* and *screen name*. Attributes such as the *user name* commonly encode relevant information about the gender of the user. Previous studies show that the 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 (Bechar-Israeli, 1995; Calvert et al., 2003; Stopczynski et al., 2014).

The main contributions are two-fold: Firstly, we assess the performance of the features using several supervised and unsupervised methods for a Portuguese dataset and an English dataset. Secondly, we show that the proposed features are compatible with both languages, and that results are improved when merging both datasets. We notice that using unsupervised methods, the increasing amount of data has positive impact on the results.

This method will allow the creation of extended labelled datasets in a semi-automatic fashion, based on an automatic annotation provided by the proposed features. Such extended labelled datasets will allow us to associate the textual content provided by the users with their gender and create gender models, purely based on the text contents (see Chapter 6).

4.1 Motivation

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

4.2 Datasets

Experiments performed in this chapter use an English and a Portuguese datasets of Twitter users. The English dataset (EN-users-sample-dataset) was extracted from one month of tweets collected during December of 2014, using the Twitter *streaming/sample* API. This method gives access to only about 1% of the actual public tweets[1]. We have restricted the data to geolocated tweets written in English, either from the United States or from the United Kingdom. The resulting dataset contained 296506 unique users that tweeted either from the United States or from the United Kingdom.

The Portuguese dataset (PT-users-sample-dataset) is a subset of the data described in Brogueira et al. (2014), and corresponds to a database of Portuguese users, restricted by users that have tweeted during October of 2014 in Portuguese language, and geolocated in the Portuguese mainland. This dataset contained 27227 unique users.

For the extraction of information, we developed API clients to connect to:

- Twiter Streaming API[2]: English users information;
- Brogueira et al. (2014) API[3]: Portuguese users' information;

[1] Limit on streaming tweets. https://dev.twitter.com/discussions/6789. (Visited on 21/02/2015).
[2] More information: https://dev.twitter.com/streaming/overview
[3] API available in the private network of INESC-ID: https://www.l2f.inesc-id.pt/.

4.3 Names dictionaries

In order to automatically associate names that can be found in the user's profile with the corresponding gender, we have compiled a dictionary of English names and a dictionary of Portuguese names. Both dictionaries focus on names that are exclusively male or female, since unisex names can be classified as either male or female. The dictionaries contain the following information:

- name
- gender
- number of occurrences

The English names dictionary contains about 8444 names. It was compiled using the list of the most used baby names from the United States Social Security Administration Official Website[4]. The Social Security Administration database provides one file per year, ranging from 1884 to 2013. Each file contains the top 1000 names of each year. To safeguard privacy, the Social Security Administration restricts the list of names to those with at least 5 occurrences. It is important to acknowledge the following characteristics of the resultant dictionary:

- All data is from a 100% sample of the records on Social Security card applications as of the end of February 2014;
- The data is restricted to births in the United States;
- Different spellings of similar names are not combined. e.g.: Caitlin, Caitlyn, Kaitlin, Kaitlyn, Kaitlynn, Katelyn, and Katelynn are considered separate names and each has its own frequency.

For the purpose of this study, we only extracted data from 1940 and beyond, where the name occurred at least 1000 times. The dictionary is currently composed of 3304 male names and 5140 female names. Some of the observed discrepancy between male and female names is due to the different spellings of similar names not being combined.

The Portuguese names dictionary contains 1659 names. It was extracted from the corpora of Baptista et al. (2005), NPro. Their Portuguese names dictionary is based on both the extraction of names from official institution lists and from previous corpora (Rocha and Santos,

[4]Popular baby names. http://www.ssa.gov/oact/babynames/limits.html. (Visited on 21/02/2015).

Algorithm 4.1 Profile names gender feature extraction.

ExtractGenderFeatures(screen name, user name):

- Read user name and screen name
- Find dictionary names in user name and in screen name
- If no name is found:
  - Remove repeated vowels in user name and screen name
  - Find dictionary names in modified user name and modified screen name
  - If no name is found:
    * Replace user name and screen name "leet speak" characters with their equivalents
    * Find dictionary names in modified user name and modified screen name
- Add found names to found names list
- For each name:
  - Find gender features
  - For each feature:
    * Validate threshold of feature
    * Add found features to found features list
- Return found features list

2000). The extraction of names from official institution lists used mainly two sources of data: a) lists of telephone subscribers, provided by telephone companies (Consortium et al., 1995); b) lists of students enrolled in Portuguese universities. The telephone subscribers list includes additional information, namely the frequency of the names in the phone list or the information pertaining to being first name or last name. The list restricts the names to those with at least 10 occurrences. We only extracted data from first names where the name occurred at least 100 times. The Portuguese names dictionary is currently composed of 875 male names and 784 female names.

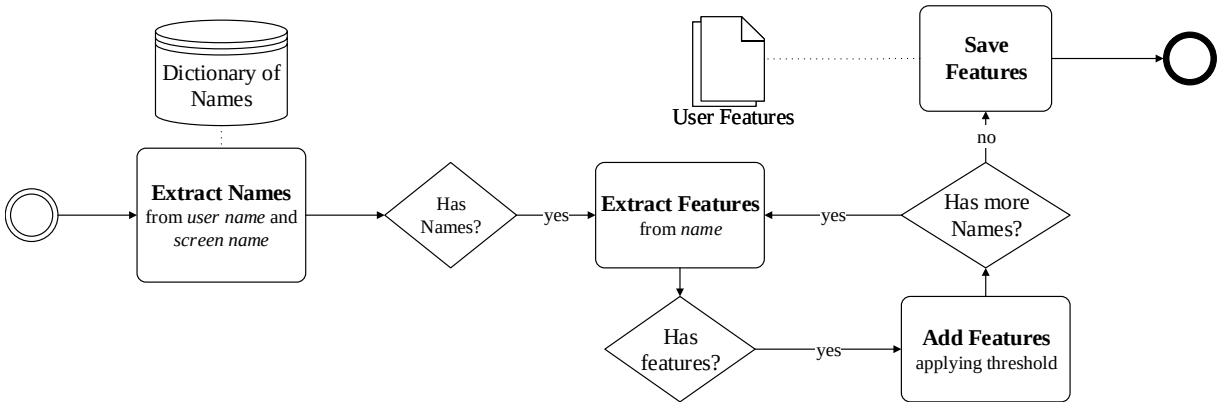


Figure 4.1: Profile names gender feature extraction diagram.

4.4 Feature extraction

For the extraction of gender features based on self-identified names with gender association, we used only the following user information: *screen name* (up to 20 characters), and *user name* (up to 15 characters). The user *description* was not considered for this task because it usually contains names related to the user's bio and preferences, and not necessarily the name of the user, making it less suitable for this task. Our gender feature extraction algorithm is summarily described in Algorithm 4.1 and illustrated in Figure 4.1.

In order to extract possible names either from the *screen name* or from the *user name*, several strategies are being applied, namely:

- Any known name found in both fields is directly extracted by consulting the names dictionary
- Names in *screen name* are found by recursively looking for names inside the string that corresponds to the entries in our dictionary of English names.

  e.g.1: Name *Ernest* in screen name *ernest_hemingway*

  e.g.2: Name *Dolan* in screen name *dodoLand77*.

- The *user name* is split into separate words using regular expressions in order to identify individual names.

  e.g.: Name *Ernest* in user name *Ernest Hemingway*.

- When no regular names are found, we assert if the *user name* or the *screen name* contain repeated vowels. If so, we remove the repeated vowels and look for names inside the modified user name and in the modified screen name words.

Leet	Character
3	e
1	l
0	o
7	t
4	a
6	g
$	s

Table 4.1: Leet Speak replacements.

e.g.: Name *Ernest* in screen name *erneeeest_hemingway*.

- Prior research shows that the choice of Twitter *user name* and *screen name* includes various stylistic forms used by internet users. Emoticons, acronyms and "leet speak" are proliferating (Corney, 2003). Leet speak is a style of writing where characters are replaced with numbers or characters which result in a similar appearance.

  E.g.: Name *Ernest* in screen name *3rn3st_hemingway*.

  When no names are found even after eliminating repeated vowels, we assert if the *user name* or the *screen name* contain leet speak characters. If so, we replace the character for the equivalent character and look for names inside the modified user name and in the modified screen name (Table 4.1).

Our model consists of 192 features. The features are prefixed with "u_" when found inside the *user name* and prefixed with "s_" when found in the *screen name*. The features are suffixed with the corresponding male (_m) or female (_f) gender (e.g. "u_name_exists_*m*"). When the features are found using modifications, such as for example, leet processing, the information is suffixed to the feature (e.g. "u_name_exists_*leet*_m").

After finding one or more names in the *user name* or *screen name*, we extract the applicable features from each name by evaluating the attributes "Case", "Boundaries" and "Position". Each attribute increases the feature granularity.

- **Case**: the case of a name has more relevance for names found in the screen name. For example,

  - user name: *ernest* hemingway (case is not relevant)
  - screen name: im*Ernest*Hemingway77 (case is relevant to separate names)

  When the case is relevant, the feature is stored as "name_exists_and_case_m"

Feature	Threshold
name_exists	5
name_exists_and_case	4
name_correct_end_separation	5
name_correct_end_separation_and_case	4
name_correct_beginning_separation	5
name_correct_beginning_separation_and_case	4
name_correct_separation	3
name_correct_separation_and_case	2
name_beginning_no_separation	4
name_beginning_no_separation_and_case	5
name_beginning_with_separation	3
name_beginning_with_separation_and_case	2

Table 4.2: A selection of existing profile name gender features.

- **Boundaries**: indicates if individual words in screen name are properly bounded, i.e., if they start with a capital letter, end with lower case and are not followed by a lower case (they can be followed by a number). Boundaries are found using regular expressions in the *screen name*; partial names are ignored.

  e.g.: EmmaRichter13; "emma" has correct boundaries while "mari" does not.

  e.g.: screen name im*Ernest*Hemingway77 has the feature "correct_end_separation_and_case_m"

- **Position**: indicates the position of the name within the *user name*.

  e.g.: user name: *ernest* hemingway has the feature "name_beginning_m"

Each of the 192 features has an associated length threshold. If the length of the extracted name is smaller than the threshold, the feature is discarded. The threshold of each feature was fine-tuned based on Logistic Regression experiments. Features with higher granularity typically have lower thresholds. Table 4.2 shows a selection of features and the corresponding thresholds. E.g., Consider *screen name "jill_gaines"*. Three names are extracted from this *screen name*, "aine", "ines" and "jill". Feature *name_exists* has threshold 5 and is therefore discarded when associated with name *"*aine*"*. Feature *name_correct_end_separation* has threshold 5 and is therefore discarded when associated with name *"*ines*"*. Feature *name_beginning_separation_f* has threshold 3 and is considered when associated with name *"*jill*"*. We can observe that all names have the same length (four characters), but the feature *name_beginning_separation_f* has a lesser threshold, since the name is at the beginning of the name.

The proposed method extracts all applicable features for each name. Consider the name "Ernest Hemingway" as an example:

feature extraction:

"name_exists_m"
"name_exists_and_case_m"
"name_correct_end_separation_and_case_m"
"name_beginning_no_separation_and_case_m"
"name_beginning_with_separation_and_case_m"

4.5 Labelled data

In order to perform the evaluation of the proposed features, we manually labelled with gender information a random selection of both Portuguese and English users. Each one of the users was associated with the corresponding gender. It is worth noting that we were labelling users to test our profile name features, meaning all users being verified had to have at least one name in their profile information. From the dataset of English users, 243522 users (82%) triggered at least one gender feature. Such value decreases significantly for the Portuguese dataset, where 15828 users (58%) triggered our proposed gender features.

We developed a Python routine to create the list of users to label with its information and profile URL (Uniform Resource Locator). We used Microsoft Excel 2013 to gather the results of the manual labelling.

Previous studies reveal that the most commonly used method to obtain a labelled datasets is through the gender/name association using the Twitter profile information (*user name* and *screen name*) (Rao et al., 2010; Pennacchiotti and Popescu, 2011; Al Zamal et al., 2012). In the research of Ciot et al. (2013), the gender is identified using the profile picture associated with the user account. The study of Burger et al. (2011) complementarily examines blog sites (found in the URL field in their profile) to label users with gender. In fact, the research of Huffaker (2004) has found convenient to verify blogs, because those blogs have profile pages with explicit gender attributes.

We have combined the three approaches and created our labelled subsets using the following method: users were manually analyzed, by validating: i) their *user name*/*screen name*, ii) their profile picture, iii) if they were human individuals, iv) possible associated blogging websites.

	English			Portuguese		
	Male	**Female**	**Total**	**Male**	**Female**	**Total**
#Users	330	418	748	249	467	716

Table 4.3: Profile names gender feature labelled data.

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

Table 4.4: Features extracted from each profile and their properties.

1. Firstly, we looked for names both in the *user name* and in the *screen name* of the profile. If the name was inconclusive, the user was discarded.

2. Secondly, we analyzed the profile picture of the user. If the picture did not correspond with the gender of the names found, the user was discarded. Users without photography or with celebrity-based pictures were discarded as well.

3. Thirdly, we assured that the author of the profile was not a bot. Previous findings suggest that about 7% of tweet profiles are non-human spam bots (Finger, 2015). We analyzed the volume of tweets per day, high number of following vs low number of followers avoiding such users. We discarded users that tweeted using the Twitter API, since people tend to tweet from the web or mobile.

4. Finally, if the user had blogging sites associated to their profile, we followed those URLs and validated the data found with their profile.

The English labelled dataset (EN-labelled-users-sample-dataset) has 748 users: 330 male users and 418 female users. The Portuguese labelled dataset (PT-labelled-users-sample-dataset) has 716 users: 249 male users and 467 female users. The majority of the users are female, which is consistent with the work of Heil and Piskorski (2009) that performed a study of correlation between name and gender, and estimates that 55% of Twitter users are female. Table 4.3 shows the distribution of the labelled datasets. Table 4.4 shows the number of features that can be extracted from the manually labelled subset as well as statistics for the extracted names in each one of the profile attributes. For English we observe more occurrences of features in *user names* (63% against 37% in *screen names*). The frequency of "Leet speak" is consistent with the general features distribution. As expected, repeated vowels occur more in *screen names* because

they must be unique for all Twitter users, unlike *user names* that impose no restrictions to their content. For Portuguese we observe more occurrences of features in *screen names* (57% vs 47% in user names). Repeated vowels related features occur more frequently in Portuguese *screen names*. The English data reveals that names in *screen name* are more unreliable. That is due to the *screen name* being a unique string without spaces, which leads to a higher uncertainty when extracting possible names. Names in *screen name* are more unreliable in English users than in Portuguese users (73% versus 16%). A similar discrepancy can be found in Chen et al. (2006), that achieved better results with Portuguese and French than with English and German, when identifying language origin of names using trigrams of letters.

4.6 Setup

For all our supervised experiments and for the unsupervised experiments with kMeans clustering, we used WEKA Explorer 3.6[5] (Waikato Environment for Knowledge Analysis). WEKA is 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 (Hall et al., 2009). We developed routines in Python, using the fuzzy logic module from the scikit-learn toolkit[6] to perform our experiments with Fuzzy c-Means clustering. To perform our experiences in the WEKA Explorer, we created datasets in the ARFF (Attribute-Relation File Format) file format. This allowed us to use the same data to perform different machine learning schemes. The file format defines the dataset with a relation composed of attributes. The ARFF header defines the name of the relation and the attributes (name and data type). The ARFF data section stores the records of the dataset, in this case, the features from each Twitter user. An ARFF file example:

[5] http://www.cs.waikato.ac.nz/ml/weka/
[6] https://github.com/scikit-fuzzy/scikit-fuzzy

```
@RELATION  en_sample_users

@ATTRIBUTE user_id STRING
@ATTRIBUTE feature_1 NUMERIC
@ATTRIBUTE feature_2 NUMERIC
@ATTRIBUTE feature_3 NUMERIC
@ATTRIBUTE class{M,F}

@DATA
'199038621',0,1,0,M
'149576071',0,1,0,M
'1154186437',1,0,1,F
'193866394',0,1,0,M
```

4.7 Evaluation metrics

In this section, we look at the metrics used to determine how good a classifier performs. These metrics will be used in all experiments of this study. Typically there are four key concepts: Precision, Recall, F-Measure and Accuracy. Before the formulas are presented, it is important to grasp the statistical definitions that constitute those formulas, within the scope of Twitter gender classification:

1. True Positive (TP): This means that a user has been correctly identified as belonging to that gender;
2. False Positive (FP): This means that a user not belonging to a given gender, has been incorrectly identified as belonging to that gender;
3. True Negative (TN): This means that a user not belonging to a given gender, has been correctly identified as not belonging to that gender;
4. False Negative (FN): This means that a user belonging to a given gender, has been incorrectly identified as not belonging to that gender;

With this in mind, the definition of the metrics are:

Precision - measures the percentage of instances classified into a particular class that were correctly classified.

$Precision = \frac{\#TP}{\#TP+\#FP}$

Recall - measures the percentage of a class that was classified correctly.

$Recall = \frac{\#TP}{\#TP+\#FN}$

F-Measure - computes the score as a weighted harmonic mean of the precision and recall. The best score is 1 and the worst is 0.

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

Accuracy - measures the percentage of correctly classified instances.

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

4.8 Experiments and results

This section describes the results obtained on the English and Portuguese datasets, and the dataset containing both English and Portuguese users, when applying supervised and unsupervised approaches based on the proposed features.

The supervised methods include: Multinomial Naive Bayes (MNB) (McCallum et al., 1998), a variant of Naive Bayes, Logistic Regression (Le Cessie and Van Houwelingen, 1992), and Support Vector Machines (SVM) (Platt et al., 1999; Keerthi et al., 2001). The unsupervised methods include Fuzzy c-Means clustering (FCM) (Bezdek et al., 1984) and k-means (MacQueen, 1967). The fuzzy logic module from the scikit-learn toolkit[7] was used for implementing FCM, and all the other methods were applied using Weka Explorer.

While the supervised based methods use labelled data to build a model, that is not the case of unsupervised methods, which group unlabelled data into clusters. For that reason, we will first describe experiments using labelled data only, and then will extend the analysis to all the data, but restricting the experiments to unsupervised methods only. Experiments using supervised methods use the labelled data for training and with a 5-fold cross-validation. Experiments using unsupervised methods use all data for creating two different clusters, the labelled data was used for validation, and each cluster was assigned to the class with more elements from that cluster. In terms of setup, k-means was set to use the Euclidean distance, centroids are computed as a mean, and the seed was set to 10. In order to use the FCM clustering algorithm, the data has been converted into a matrix of binary values, and we have used 1000 iterations, and the Euclidean distance. All experiments consider binary features.

Results achieved with each one of the methods are summarized in Table 4.5. The first 3 rows show the performance for supervised methods. Results from the last two columns were

[7]https://github.com/scikit-fuzzy/scikit-fuzzy

	English		Portuguese		English + Portuguese	
	Accuracy	Kappa	Accuracy	Kappa	Accuracy	Kappa
Logistic Regression	93.7%	0.872	97.6%	0.951	96.3%	0.927
Multinomial Naive Bayes	**97.2%**	**0.943**	**98.3%**	**0.964**	**97.9%**	**0.966**
Support Vector Machines	96.4%	0.931	97.8%	0.952	97.4%	0.950
kMeans clustering	67.3%		70.1%		67.8%	
Fuzzy c-Means	**96.0%**		**94.4%**		**96.4%**	

Table 4.5: Gender classification results for supervised and unsupervised methods.

achieved by combining both the English and the Portuguese labelled subsets. MNB achieved the best performance for both languages, and achieves even better performance for the merged subset of users, achieving about 98% accuracy, proving that datasets can be combined and that features are compatible with the two languages. The achieved performance suggests that the proposed features can be suitable to discriminate the user's gender for both languages. The last two rows of the table summarizes the performance for unsupervised methods. FCM obtains the correct gender for about 96% of the English users and about 94% of the Portuguese users when all the data is used. k-means achieves a much lower performance for both languages. The last column of the table shows the results when English and Portuguese data are combined. With such dataset, FCM achieves the best results so far, outperforming individual results obtained for each language.

Our proposed features compare well with the performance achieved by other state-of-the art research. For example, Burger et al. (2011) uses the winnow algorithm with ngrams extracted from the user's full name and obtain 89.1% accuracy for gender detection when using only *user name* and *screen name*.

4.9 Assessing the impact of amount of data

We have performed additional experiments in order to assess the impact of using increasing amounts of data. Figure 4.2 shows the impact of the amount of data on the performance of FCM, revealing that it has positive impact until reaching the 50k users. Above that threshold, the accuracy tends to remain stable, which may be due to our relatively restricted set of users.

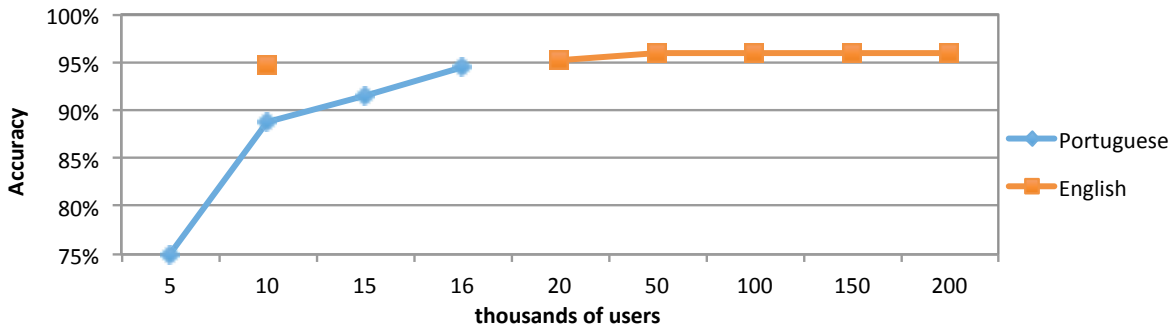


Figure 4.2: Impact of the amount of data on the performance, for Portuguese and English.

4.10 Discussion

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

Towards extended labelled datasets

5

We can only see a short distance ahead, but we can see plenty there that needs to be done.

Alan Turing

This Chapter proposes a new method for creating extended labelled datasets in a semi-automatic fashion, using the features obtained in Chapter 4. Such extended labelled datasets will allow to associate the textual content provided by the users with their gender and create gender models, purely based on the text contents.

5.1 Motivation

The creation of Twitter corpora is not new and researchers have built both English (Petrović et al., 2010; McCreadie et al., 2012) and Portuguese (Brogueira et al., 2014) Twitter users corpora. The problem is that those corpora are not labelled with user attributes like gender or age. In order to create supervised gender models, a labelled dataset is needed. Previous studies reveal this task to be demanding, labor-intensive and not reusable. Rao et al. (2010) manually annotated 1000 profiles through the gender/name association using the Twitter profile information (*user name* and *screen name*). Likewise, Liu et al. (2012); Deitrick et al. (2012); Miller et al. (2012); Ciot et al. (2013); Pennacchiotti and Popescu (2011); Al Zamal et al. (2012); Kokkos and Tzouramanis (2014); Ugheoke (2014); Nguyen et al. (2014)[1] and us, in Chapter 4, manually labelled users to produce datasets, observing either *user name*, *screen name*, profile picture, tweets or a combination of those attributes.

[1] Nguyen et al. (2014) and us, in Chapter 4, also verified information available in social media profiles such as Facebook and Linkedin and associated blogging websites, when provided by Twitter users.

Work	Users	Tweets	Language	Geography
Rao et al. (2010)	1000	405k	English	India
Burger et al. (2011)	183729	4.1M	Several	
Liu et al. (2012)	400	N/A	English	Canada
Bamman et al. (2012)	14464	9.2M	English	United States
Deitrick et al. (2012)	N/A	3031	English	
Fink et al. (2012)	11155	18.5M	English	Nigerian
Miller et al. (2012)	3000	N/A	English	
Al Zamal et al. (2012)	400	N/A	English	Canada
Bergsma et al. (2013)	This study used the dataset from Burger et al. (2011)			
Liu and Ruths (2013)	8000	8M	English	
Ciot et al. (2013)	943	N/A	French	
	3237	N/A	Indonesian	
	3609	N/A	Turkish	
	829	N/A	Japanese	
Kokkos and Tzouramanis (2014)	N/A	10000	English	
Ugheoke (2014)	1000	N/A	English	
Halteren and Speerstra (2014)	600	N/A	Dutch	
Nguyen et al. (2014)	3000	N/A	Dutch	
Van Zegbroeck (2014)	8791	N/A	Flemish	
Our study reported in Chapter 4	748		English	
	716		Portuguese	
Our study reported in Chapter 5	65073	6.5M	English	
	57705	5.8M	Portuguese	

Table 5.1: Twitter labelled datasets of previous works.

Burger et al. (2011) followed the blogging website links available in the profile of Twitter users, and extracted the gender from the corresponding profiles. To evaluate the accuracy of their method, they randomly selected 1000 Twitter users and manually validated them. Only 15% of the sample had explicit gender information. In this case, filtering only Twitter users with blogs may bias the dataset, but also filters bots and spammers.

Liu and Ruths (2013) labelled their data using the Amazon Mechanical Turk platform. A platform developed for the distribution of tasks to human workers. Each human intelligence task (HIT) is performed by a person in exchange for a small payment. The reliability of such method is uncertain, even when the same task is performed by more than one person. In the study from Burger et al. (2011), the accuracy of Amazon Mechanical Turk human gender classification was of 68.7%, when averaged across works. Table 5.1 shows the labelled datasets used in previous works. A general observation is that most of the studies use small labelled datasets, except for Burger et al. (2011) that uses a large number of users. It also uses several languages, but notice that English represent 66.7% of the users, Portuguese 14.4% and Spanish 6%.

It is our contention that this task can be improved by using the features proposed in Chapter 4 in a semi-automatic fashion.

5.2 Data

Experiments performed in this chapter use an English and a Portuguese datasets of Twitter users. In both datasets, we retrieved only the last 100 tweets of each user.

The English dataset (EN-users-full-dataset) was extracted from one year of tweets collected since January till December of 2014, using the Twitter *streaming/sample* API, limited to only about 1% of the actual public tweets. We have restricted the data to English language, to users with at least 100 tweets and to only 100000 users.

The Portuguese dataset (PT-users-full-dataset) is the full dataset of the data described in Brogueira et al. (2014), 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. This dataset contained 105490 unique users.

5.3 Proposed approach

In this section, we propose an approach to create extended labelled datasets in a semi-automatic fashion. To do so, we must complete the following steps: i) extract data from twitter; ii) filter the dataset; iii) create a gender classification model; iv) classify dataset users; v) validate the quality of the data; vi) enrich the dataset (optional); vii) create data subsets. After creating the datasets we partially validated the data. The following subsections describe these steps in more detail.

5.3.1 Data extraction and filtering

The data extraction and filtering used for the creation of both the Portuguese and the English datasets are explained in Section 5.2.

Algorithm 5.1 Extended labelled datasets creation.

ExtendedDatasetCreation(screen name, user name):

- Extract proposed features from *user name* and *screen name*
- If features found:
  - Classifies user with MNB gender classification model
  - If confidence >= 95:
    * Adds user profile information and last 100 tweets to extended dataset

	No Features		1 to 10 Features		More than 10 features	
Dataset	# Users	%	# Users	%	# Users	%
Portuguese	44559	42%	57440	55%	3451	3%
English	27110	27%	65559	66%	7331	7%

Table 5.2: Automatic gender feature extraction results.

5.3.2 Gender classification model

A Multinomial Naive Bayes approach was used to build the gender classification models, based on the features used in Chapter 4. This choice was motivated by the experiments performed, where Multinomial Naive Bayes turned out to give the best performance. Two different models were created, for English and Portuguese, and will be made available online for future usage.

5.3.3 Dataset classification

For the classification of the dataset, we used only the following user information: *screen name* (up to 20 characters), and *user name* (up to 15 characters). Our suggested gender classification algorithm is summarily described in Algorithm 5.1.

We ran the algorithm in our Portuguese and English datasets. In the English dataset, only 27.1k (27%) of the users had no profile name feature. 65.5k users had 1 to 10 features and 7.3k users had more than 10 features. In the Portuguese dataset, 44.5k users had no feature (42%), a higher percentage when comparing to the results obtained in the English dataset. 57.4k users had 1 to 10 features and 3.4k users had more than 10 features. Results achieved with each dataset are summarized in Table 5.2. Figure 5.1 details the distribution of features per user, revealing that the distribution is identical between Portuguese and English.

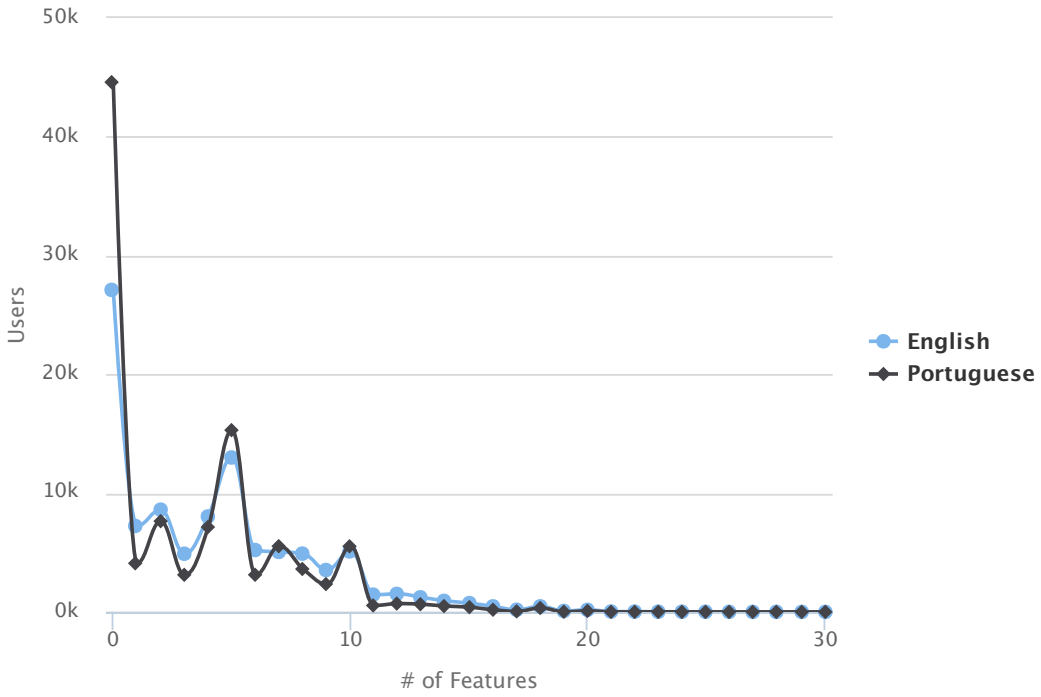


Figure 5.1: Automatic Gender Classification - Features per users.

	Features							
Dataset	None		User name		Screen name		Both	
Portuguese	44599	42%	17776	18%	18443	17%	24672	23%
English	27110	27%	20845	21%	20580	21%	31465	31%

Table 5.3: Automatic gender feature extraction results per attribute.

We further analyzed the distribution of features in the *screen name* and the *user name*. Since the *screen name* does not contain spaces, there was the possibility that the English dataset might obtain more users with features due to *screen name* features, where names might be the result of parts of concatenated words.

In the English dataset, 41.4k users had only features in one of the attributes (*screen name* or *user name*). The distribution of these was even, 20.5k users had features only in the *screen name* and 20.8k users had features only in the *user name*. 31.4k users had features in both fields.

In the Portuguese dataset, the results were similar. 36.2k users had only features in one of the attributes (*screen name* or *user name*). The distribution of these was also even, 18.4k users had features only in the *screen name* and 17.7k users had features only in the *user name*. 24.6k users had features in both fields. Table 5.3 shows the results obtained.

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

Table 5.4: Some of the gender indicative words.

5.3.4 Validate data quality

During the automatic gender classification stage, users with no features or with features, but classified with a confidence inferior to 95% were already discarded. To further improve the quality of the data, we manually validated a subset performing the two following tasks:

1. randomly select a sample of the labeled dataset to manually validate and correct;
2. select a sample of the labelled dataset by searching for gender related words in the users' descriptions to manually validate and correct.

Concerning the second task, Table 5.4 describes some of the words more informative about the gender. Some of these words are associated to the opposite gender when preceded by possessive determiners (e.g.: "my husband" is considered female[2], while "husband" is male). This second task turns out to be biased, since the probability of finding wrong classification is higher, but certainly improves the quality of the dataset.

5.3.5 Enriching the dataset

In order to enhance the datasets further, we added two new attributes for each user: gender recognition from profile picture, and detailed geographical information based on the last known location. The first attribute provides useful information for the gender classification, while the second attribute is relevant for tackling region specific phenomena.

5.3.5.1 Gender based on the profile picture

Something that was not seen in previous work, is the use of the gender attribute extracted from the profile picture. However, the profile picture might contain clues regarding the gender

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

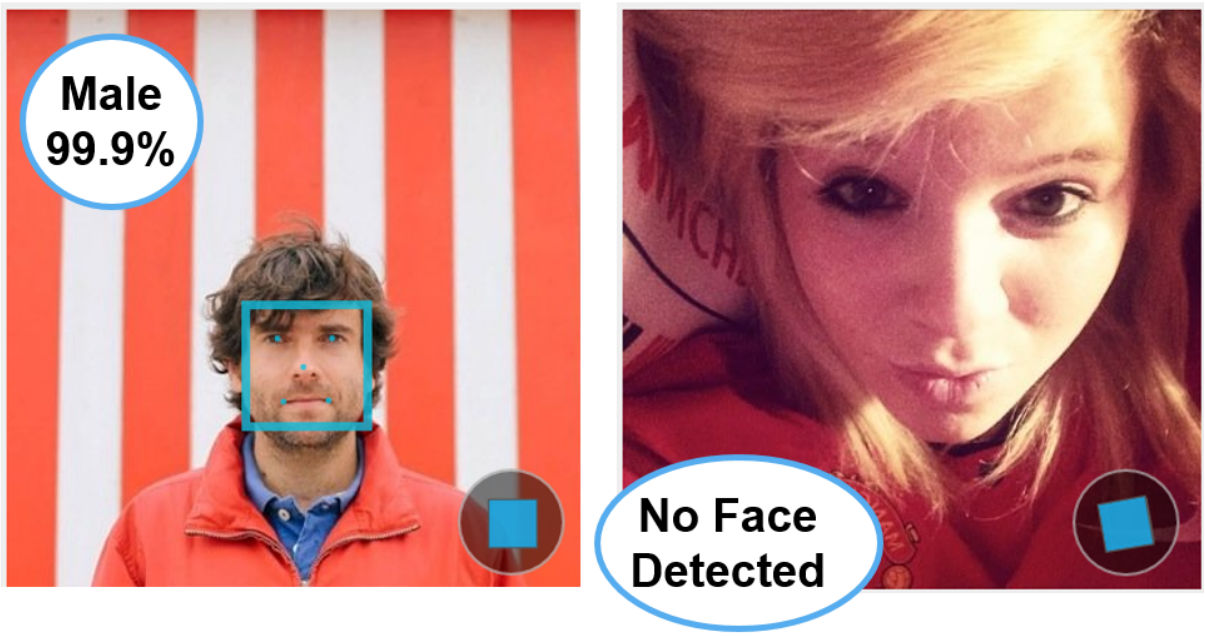


Figure 5.2: Face++ gender detection examples.

of the user.

Face++[3]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. Figure 5.2 illustrates the usage of Face++, where the first picture was correctly classified.

Limitations: Since our datasets contain past data (from 2014), some of the users have already changed their profile picture. More, some of the pictures did not contain faces, or the Face++ was not able to detect the face in the picture. Face++ was unable to identify the face in the second picture of Figure 5.2.

Table 5.5 summarizes the gender data retrieved from the Face++ API. 54% of the English users and 44% of the Portuguese users have changed their profile picture since 2014. From the users with an existing profile picture, for 36% in both datasets no face was detected. In the English dataset, more male users (34%) than female users (29%) have a profile picture with a face. In the Portuguese dataset, the opposite occurs, more female users (35%) than male users (30%) have a profile picture with a face.

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

	English dataset			Portuguese dataset		
	# Users	%	%	# Users	%	%
Image Unavailable	31076	54%		28605	44%	
No Face detected	9777	17%	36%	12995	20%	36%
Male	9156	16%	34%	10805	17%	30%
Female	7857	14%	29%	12649	19%	35%

Table 5.5: Face++ gender data retrieved.

5.3.5.2 Geographical location

People might write differently according to their location. Twitter provides geolocation in each tweet if the user allows geolocation. For a better use of this metadata, we added geographical information to our datasets. We took different approaches depending on the dataset.

In the Portuguese dataset, we added a feature with the district of the location. We extracted the last location from the user and searched for a city or district. Examples of geolocation:

Lisboa, Portugal

Paços de Ferreira, **Porto**

Vila Nova de Gaia, Portugal

From the above example, in bold, we distinguish districts, in italic, cities or locations. After finding cities, we mapped them to the corresponding district. In the case of the Portuguese archipelagos, we aggregated each location in its archipelago, Madeira and Azores. Finally, we added the district information to each user. Table 5.6 shows the distribution of users by district. Figure 5.3 shows a geographical distribution of the Portuguese labelled users.

The English dataset contains tweets in English from more than 200 countries. To add state or district information for each country would be almost impossible and in most cases unnecessary, since for more than 100 countries the dataset contains less than 10 users. From the entire labelled dataset, 78% users' last geographical location was the United States and 11% the United Kingdom.

For the United State users, we added the information regarding the location's state. We extracted the last location from the users and searched for a city or state. Examples of geolocations:

New York, **NY**

St. James, **NY**

New York, US

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

Table 5.6: Portuguese users by district and gender.

Figure 5.3: Portuguese labelled users per district.

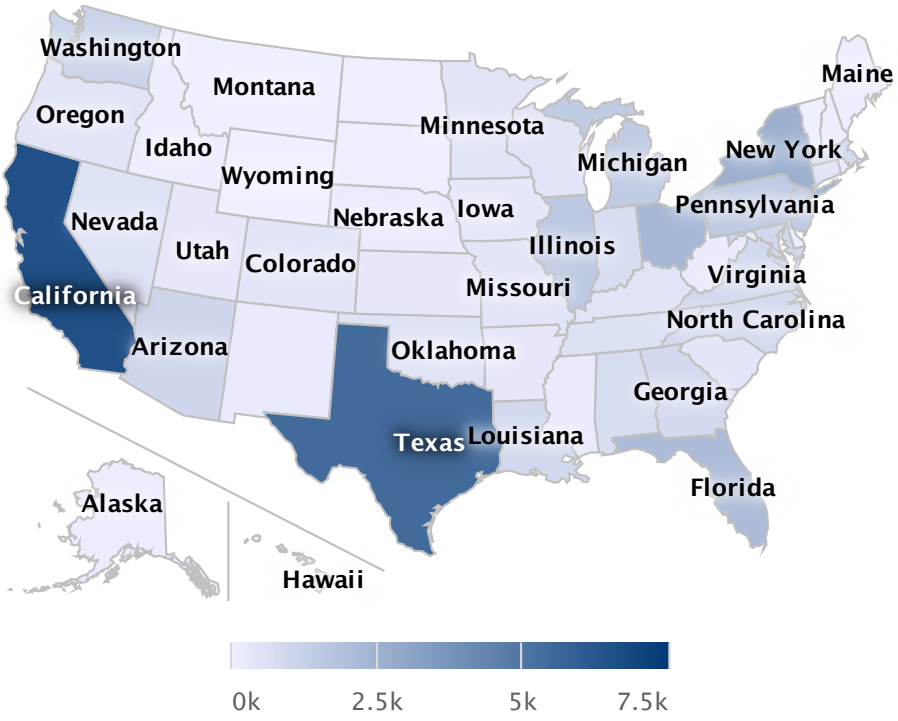


Figure 5.4: United States labelled users per state.

New Jersey, USA

From the above example, in bold, we distinguish states, in italic, cities or locations. For tweets geolocated in the United States, most of the times, Twitter provides the state code (from the standard INCITS 38[4]). When the code was not found, we extracted the location and mapped to the corresponding state code. Figure 5.4 shows the distribution of the United States labelled users per state.

For the United Kingdom labelled users, the distinction added was the country: Scotland, Northern Ireland, England and Wales. We extracted the last location from the users and searched for a city, a state or a country. Examples of geolocations:

North East, United Kingdom

Westminster, *London*

Cardiff, **Wales**

From the above example, in bold, we distinguish countries, in italic, cities or locations. When the country was not found, we extracted the location and mapped to the corresponding country. Figure 5.5 shows the distribution of the United Kingdom labelled users per country.

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

Figure 5.5: United Kingdom labelled users per country.

Dataset	Total	Train	Validation	Test
English	65073	39043	13015	13015
Portuguese	57705	34625	11540	11540

Table 5.7: Description of obtained semi-automatic gender labelled datasets.

5.3.6 Creating data subsets

As previously mentioned, only users classified with a confidence above or equal 95% were kept in our datasets. The resulting datasets were partitioned in 3 subsets:

- **Train**: to train new models;
- **Validation**: to train and improve models, avoiding probable overfitting problems;
- **Test**: to assess the performance of the model.

Table 5.7 summarizes the gender labelled datasets and the corresponding subsets. We partitioned the datasets with the following ratio: Train subset, 60% of the users, validation and test subsets, 20% of the users.

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

Table 5.8: Manual validation of automatic gender classification.

5.3.7 Data validation

To ensure the quality of the proposed method, we performed a manual validation of a sample of data. We randomly chose about 3k users from each labelled dataset and validated both the Twitter profile content and the blogging sites (when available).

We looked for names both in the *user name* and in the *screen name* of the profile, analyzed the profile picture of the user and, if the user had blogging sites associated to their profile, we followed those URLs and cross validated the data found with their gender classification. At the end of this process, we concluded that most of the incorrect classifications in the datasets were due to:

- Gender incorrectly assigned;
- Twitter profile was not of a person;
- User was transsexual;
- Profile was removed and the manual validation was impossible to perform.

Table 5.8 summarizes the results obtained. In the English dataset we detected 9% of incorrect automatic gender classification. It is a high percentage of error, considering we had 97.9% accuracy in Chapter 4, using a smaller but manually labeled dataset. In the Portuguese dataset we detected only 3% incorrect automatic gender classification. The difference in the accuracy might be related to the higher percentage of English users with features, probably due to noise found in the name attributes. Also, Portuguese language has a construction of names with more clues to gender than English.

Observing the profiles incorrectly classified, it is possible to notice that female names represent a higher percentage in both datasets. In the English dataset, they represent 68%, a similar percentage when compared with the variation of the sample. In the Portuguese dataset the difference is noticeable. Female users represent 82% of the errors, even though the random sample contained only 58% of female users.

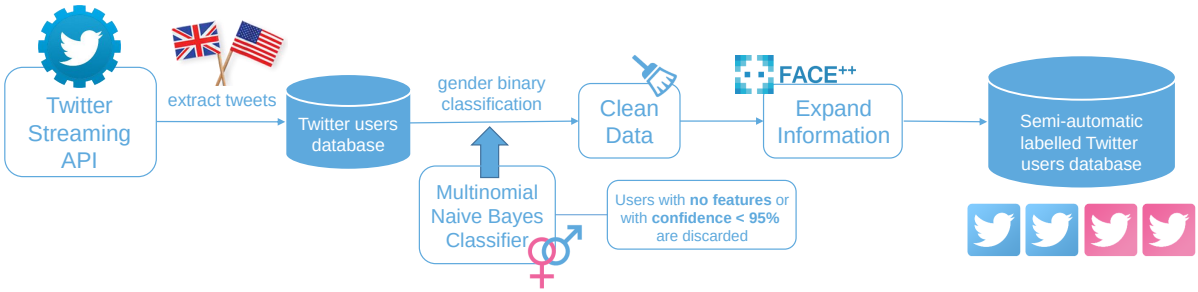


Figure 5.6: Semi-automatic gender labelled dataset creation diagram.

5.4 Conclusion

In this chapter we proposed an approach for the creation of extended labelled datasets in a semi-automatic fashion. This method allows the creation of gender labelled twitter users datasets in a inexpensive and reusable way. Our labelled datasets are only surpassed in size by Burger et al. (2011), but with less effort and limited resources. Figure 5.6 illustrates an example of semi-automatic gender labelled dataset creation, filtering for English Twitter users' geolocated in the United Kingdom and in the United States.

The proposed method has still the following limitations:

- Twitter users might not use their real names. Therefore, the reliability of self-declared names is uncertain (e.g.: a male user can have a female gender associated *user name*).
- Our method does not filter for profiles of companies and other organizations;
- Twitter metadata might be incorrect. For example, a tweet identified by Twitter as being written in Portuguese may be written in a different language;
- The performance over English data is not as good, when compared to Portuguese.

Combined gender classification

6

By this art you may contemplate the variation of the twenty-three letters.

Robert Burton, *The Anatomy of Melancholy, part 2, sect. II, mem. IV*

In Chapter 4, we explored gender classification using only user's profile information. We successfully experimented both supervised and unsupervised methods, with a dataset of manually labelled Twitter users. In Chapter 5, we created an extended labelled dataset using the gender model created in Chapter 4. In this chapter, we propose a method for gender detection using a combined classification. We will use the Portuguese and English labelled datasets from Chapter 5 for all the experiments of this chapter. In Section 6.3 we describe the different attributes chosen to help predict the gender of a user. The features, based both on the users' content and profile information, are distributed in the following groups: *user name* and *screen name*, *description*, tweet content, profile picture and social network. Finally, in Section 6.4, we describe the classifiers used for each group of features and the combined classifier. We report the performed experiments and discuss the results obtained.

6.1 Motivation

Instead of using 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. This approach provides two advantages: 1) enables the choice of the best classifier for each group of features; 2) improves the accuracy obtained when comparing to the separate classifiers. Figure 6.1 shows the combined classifier. It receives five outputs sent from the separate classifiers and uses them as inputs to generate a new prediction.

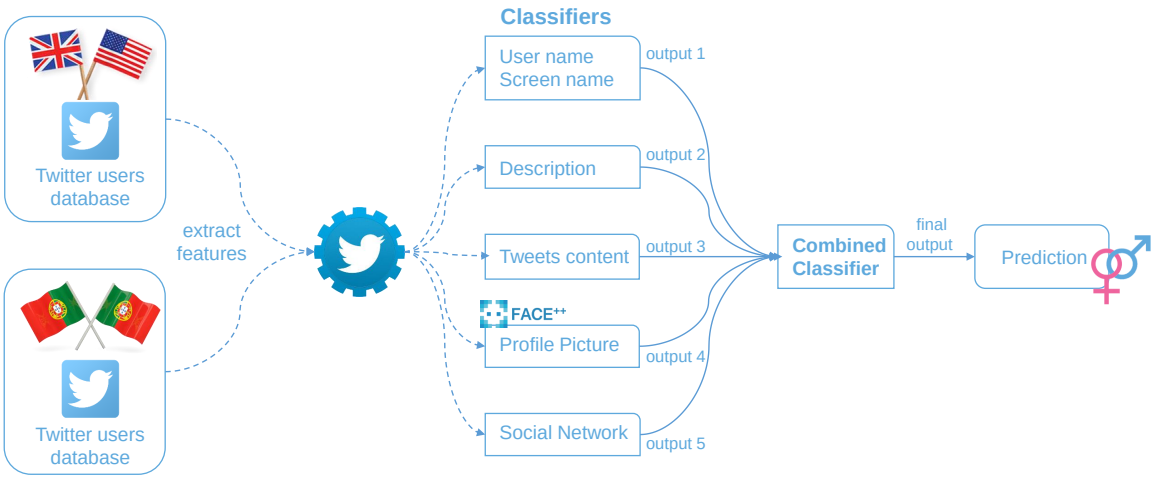


Figure 6.1: Combined classifier: output of each classifier is input for the combined classifier.

Dataset	Users	Train	Validation	Test
English Users	65063	39043	13015	13015
Portuguese Users	57705	34625	11540	11540

Table 6.1: Description of gender labelled users datasets.

6.2 Datasets

A corpus of labelled Twitter users is needed to evaluate gender classifiers. For our experiments, we used both the Portuguese and English labelled datasets from Chapter 5. These datasets are used in the remainder of the chapter, unless stated otherwise. In order to be able to train and validate the classifiers, the datasets were divided in three subsets: training, development and test, with the sizes shown in Table 6.1. 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 we used the validation subset to select the final model.

6.3 Features

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

screen name, *description*, tweet content, profile picture and social network. Figure 2.2, from Chapter 2, shows the several attributes that might contain clues to infer the user gender.

All feature extraction algorithms were implemented using Python 3.4[1]. Data preprocessing and transformation routines were also developed in Python with the support of the NLTK (Natural Language Toolkit) 3.0 package[2]. NLTK provides a collection of NLP modules.

6.3.1 User name and screen name

User name and *screen name* are valuable attributes. As we stated before, 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 (Bechar-Israeli, 1995; Calvert et al., 2003; Stopczynski et al., 2014). In the study of Stopczynski et al. (2014), 92% of the inquiries stated they posted real name on social media profiles. The 192 features extracted from *user name* and *screen name* are described in Chapter 4. *User name* and *screen name* attributes might not contain any name or any distinct gender characteristic. From the dataset of English users, 82% triggered at least one feature and from the Portuguese dataset, 58% triggered features. A different approach would be to extract word and character ngrams from these attributes, as did Burger et al. (2011) and some other posterior studies.

6.3.2 Description

Users might provide clues of their gender in the description field. Having up to 160 characters, the description is optional. Table 6.2 lists some random descriptions from users of our 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;

[1] https://www.python.org
[2] http://www.nltk.org/

Gender	Dataset	Description	Tweet
Female	Portuguese	19, Moçambicana. Psicologia no ISCTE-IUL.	Ah, por favor, não se iluda. Talvez chamem você de “amor” porque esqueceram seu nome.
Female	English	I love being a mother.Enjoy every moment.	FINALLY http://t.co/NF88TgFUrq
Female	English	Sophomore • Sing • Dance • Lover • Daughter of God • Servant of the Lord	Who does that?
Female	English	19| Chill vibes only #PlayGod$™ Southern University	@KelseyAshley10 right :(I thought it was suppose to be back last month!
Male	English	Southerner	First shower, then off to the barber shop to cut my hair/beard
Male	Portuguese	Não sei, ainda ando perdido	Bora ao cinema?? XD http://fb.me/6GNvq5YvN
Male	English	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
Male	Portuguese	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

Table 6.2: Random Twitter user descriptions and tweets from labelled datasets.

- 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”.

After the preprocessing, we extract 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.

Figure 6.2: Most used words by English female and male users, respectively.

6.3.3 Content of the tweets

Features extracted from tweets content can be divided in two groups. 1) Textual ngram features, like used in Burger et al. (2011), or 2) content, style and sociolinguistic features, like emoticons, use of repeated vowels, exclamation marks or acronyms, as used in Rao et al. (2010). Table 6.2 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 Subsection 6.3.2. 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 Burger et al. (2011), 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. Figure 6.2 show the 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. Cheng et al. (2011) suggest word-based

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	

Table 6.3: Style and sociolinguistic features.

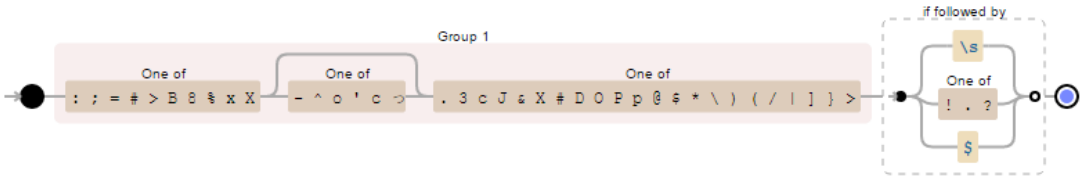


Figure 6.3: Description of the regular expression that matches smileys.

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 6.3 lists some of our features.

Features were extracted using regular expressions. E.g.: regular expression used to find smileys in text:

```
REG = r"([\:\;\=#\>B8%xX][\-\^o'c]?[\.3cJ&X#DOPp\@\$\*\\\)
    \(\/\|\]\}\>])(?=\s|[\!\.\?]|$)"
smileys = re.findall(REG, tweet)
```

Figure 6.3 explains the structure of the regular expression used to extract smileys.

For both the textual ngram features and the style and sociolinguistic features, we only used the last 100 tweets from each labelled user.

6.3.4 Profile picture feature

Profile pictures have not been used in previous studies of gender detection of Twitter users. The main reasons are: the profile picture is not mandatory; many users tend to use profile pictures of celebrities or characters from movies and TV series; the picture might 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. We use FACE++ API[3] to retrieve information regarding facial recognition of both Portuguese and English users. We invoke the API sending the URL of the profile picture and it returns the gender and the confidence. In some cases, the API does not detect any face in the picture. Subsection 5.3.5 describes the process of gathering this information.

6.3.5 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, increment the global performance. Another possible approach would be to extract attributes from the users followed by each user to infer gender, as studied by Al Zamal et al. (2012) and Bamman et al. (2012).

[3]More information: http://www.faceplusplus.com/api-overview/

6.4 Experiments and results

In this section we present the experiments and classifiers used to learn from the labelled datasets for each group of features and for the combined classifier. The purpose is to be able to classify a user outside of our labelled datasets as male or female, solely based on the combination of features extracted from profile information and tweets content. This task is known in NLP as generalization. We use the train subset of each language to train the classifiers, the validation subset to fine tune the classifiers. The test subset is only used to evaluate the accuracy of the classifiers in the end. For each user, we create a vector and assign a class, "F" for female labeled users and "M" from male labelled users. We apply supervised learning techniques, namely MNB, Logistic Regression, C4.5 decision tree and SVM. Unlike Chapter 4, we will not use unsupervised learning. For all our experiments, we used WEKA Explorer 3.6[4].

6.4.1 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 (Feldman and Sanger, 2007). 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 (Harris, 1954). 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:

[4]http://www.cs.waikato.ac.nz/ml/weka/

```
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]
])
```

The data is stored in ARFF files, as we described in Section 4.6. Each file stores the features from the labelled users of a subset.

6.4.2 Classification using user name and screen name

The results previously obtained with the *user name* and *screen name* features are described in detail in Chapter 4. The 192 features allow to infer gender when the user self-assigns a name either in the *user name* or the *screen name*. With MNB, the achieved accuracy was of 97.9% for English users and of 98.3% for Portuguese users. In Chapter 4, 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 chapter, 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.

To be able to test these features without any bias, we randomly labelled 1000 Portuguese users and 1000 English users from the datasets used in Chapter 4.

For the experiments of the *user name* and *screen name* features classifiers, we used a 5-fold cross-validation. MNB 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 Chapter 4. Though the Portuguese dataset has a higher baseline, the percentage of users with features is inferior to the English dataset, as we reported in Chapter 5. Results achieved for each of the methods are summarized in Table 6.4.

	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

Table 6.4: Gender classification results for user name and screen name features.

6.4.3 Classification using the user description

To evaluate the description features, we will use the English dataset split in three subsets as described in Chapter 5. The description field is not mandatory and from the 65063 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 in Subsection 6.3.2. 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 TF-IDF conversion and normalizing word frequencies. We applied dimensionality reduction, because the descriptions of all users are represented by thousands 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.

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 unigrams, bigrams and trigrams combined, keeping 10000 instances. Table 6.5 shows the results obtained. MNB 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

	Accuracy	Kappa	Precision	Recall	F-Measure
Baseline	51.8%				
Support Vector Machines	60.0%	0.182	63.8%	60.0%	0.5659
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
C4.5 Decision Tree	58.9%	0.164	60.5%	58.9%	0.563

Table 6.5: Gender classification results for description features of English users.

purpose, is necessary to analyze all users. These results are consistent with the work of Burger et al. (2011), where the description is the less gender indicative field.

Some of the most strong description features of English users are shown in Table 6.6. There are more informative features for the female class, being similar to previous works by Burger et al. (2011) or Van Zegbroeck (2014). Some female features are related to sentiments, as *love* or *i love*, and beauty, as *hair* or *my hair*. Male features use sports related semantic, as *game*, *team*, *win* or *lebron* or interjections, as *man* or *bro*.

6.4.4 Classification using tweets content

For the experiments using tweets content, we will use the English dataset split in three subsets as described in Chapter 5. The last 100 tweets from each user were extracted and the tweets text was preprocessed as explained in Subsection 6.3.2.

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, Table 6.7 lists the strongest ngrams by gender.

Table 6.8 shows the results obtained using the previously described parameters. Column "Time (s)" contains the time spent to build each model. SVM using unigrams achieves the highest performance, obtaining an accuracy of 73.8%. Using a combination of unigrams, bigrams and trigrams, both SVM and Logistic Regression obtain an accuracy of about 73%, but the Logistic Regression is considerably faster to build a model.

Rank	Feature	Gender	Probability
1	bro	M	0.74
2	omg	F	0.68
3	game	M	0.85
4	love	F	0.56
5	so	F	0.52
6	bc	F	0.69
7	i love	F	0.59
8	team	M	0.79
9	cute	F	0.63
10	my hair	F	0.71
11	me	F	0.51
12	mom	F	0.62
13	hair	F	0.64
14	my mom	F	0.65
15	man	M	0.89
16	win	M	0.81
17	love you	F	0.63
18	lebron	M	0.68
19	my	M	1.00
20	i m so	F	0.63

Table 6.6: Selection of the most informative description features of English users' dataset.

We applied dimensionality reduction due to the time consumed to experiment SVM based models. MNB algorithms have almost a similar performance, but is more than ten times faster. We experimented MNB using the same parameters but without feature selection. Table 6.9 shows the results. Using a combinations of unigrams, bigrams and trigrams, the performance of MNB 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 SVM 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 MNB experiences, took longer in our first experiments. Building a MNB model with unigrams and 1000 tokens lasted 119 seconds in the first experiments, but only 26 seconds in the experiments where only MNB 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 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

Rank	Female	Male
1	my hair	nigga
2	boyfriend	man
3	omg	play
4	ugh	bruh
5	cry	game
6	my mom	games
7	hair	the game
8	cute	football
9	i love you	win
10	miss you	fans
11	love you	played
12	i m so	team
13	mom	ball
14	literally	bro
15	seriously	beat
16	i miss	against
17	so much	playing
18	baby	shot
19	okay	on the
20	i hate	go

Table 6.7: Selection of the most informative textual ngram features of English users' dataset.

6.10.

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 6.11describes 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 MNB and SVM, while Logistic Regression decreased the performance. When evaluating United States subset, having 41k users, the accuracy improved in all algorithms. SVM increased almost 1%, MNB 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. Dimension-

	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 6.8: Gender classification results for textual ngram features of English users.

ality 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 6.12 shows the results of the textual ngram features in the Portuguese dataset. SVM and MNB 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 MNB 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.

6.4.5 Classification using the profile picture

To evaluate the profile picture, we use both datasets described in Chapter 5. The Twitter profile picture is extracted and sent as parameter to the Face++ API[5]. When a face is detected in the profile picture, we send the detected gender and confidence as input to the combined

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

Order	Tokens	Time (s)	Accuracy	Kappa
1	1000	26	71.7%	0.433
1	10000	30	72.8%	0.452
1	20000	33	72.4%	0.446
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	20000	224	73.1%	0.460
1-3	50000	224	73.1%	0.460
1-3	100000	259	**73.2%**	0.462

Table 6.9: Gender classification results for textual ngram features of English users using MNB.

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

Table 6.10: Gender labelled subsets of United Kingdom and United States users.

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 6.13 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.

6.4.6 About social network features

We explored the social network features described in Subsection 6.3.5. These features were not indicative of gender. We observed no differences in the social network feature values between male and female and the results were identical to the baseline accuracy. These results are consistent with the study of Rao et al. (2010). They analyzed users' network structure and communication behavior and observed the inability to infer gender from those attributes.

	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

Table 6.11: Gender classification results for textual ngram features of English users using geographical context.

	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 6.12: Gender classification results for textual ngram features of Portuguese users.

6.4.7 Combined classifier

In the previous subsections, we evaluated the separate classifiers. A summary of the results obtained is shown in Figure 6.4. 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 achieves a higher accuracy by far. 93.5% of accuracy when evaluating the last 100 tweets of each user. The English only achieves an accuracy of 73.8%,

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

Table 6.13: Gender classification results using profile picture.

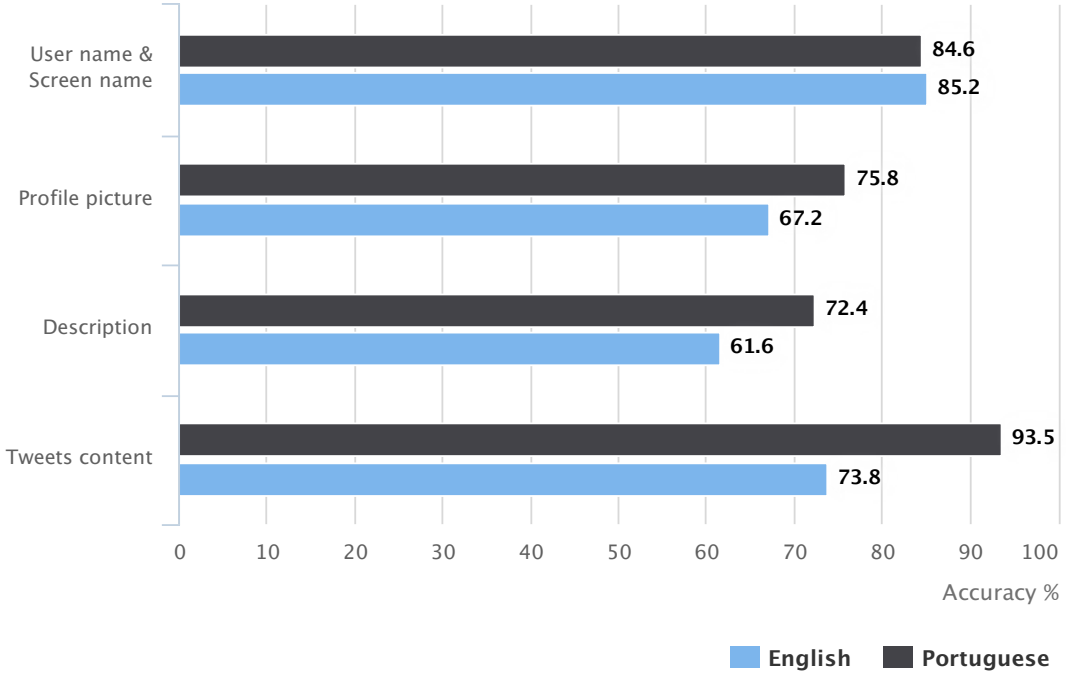


Figure 6.4: Separate classifiers' accuracy results.

which is coherent with the study of Burger et al. (2011) 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 English and Portuguese users. The datasets used to evaluate the combined classifier are described in Chapter 5. 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 SVM algorithm to evaluate the combined classifier. A number of different parameters

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

Table 6.14: Gender classification accuracy using the combined classifier.

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 6.14 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. As we seen in Chapter 4, if only considering users with a name in the profile information, an accuracy of 97.9% for English users and of 98.3% for Portuguese users were obtained.

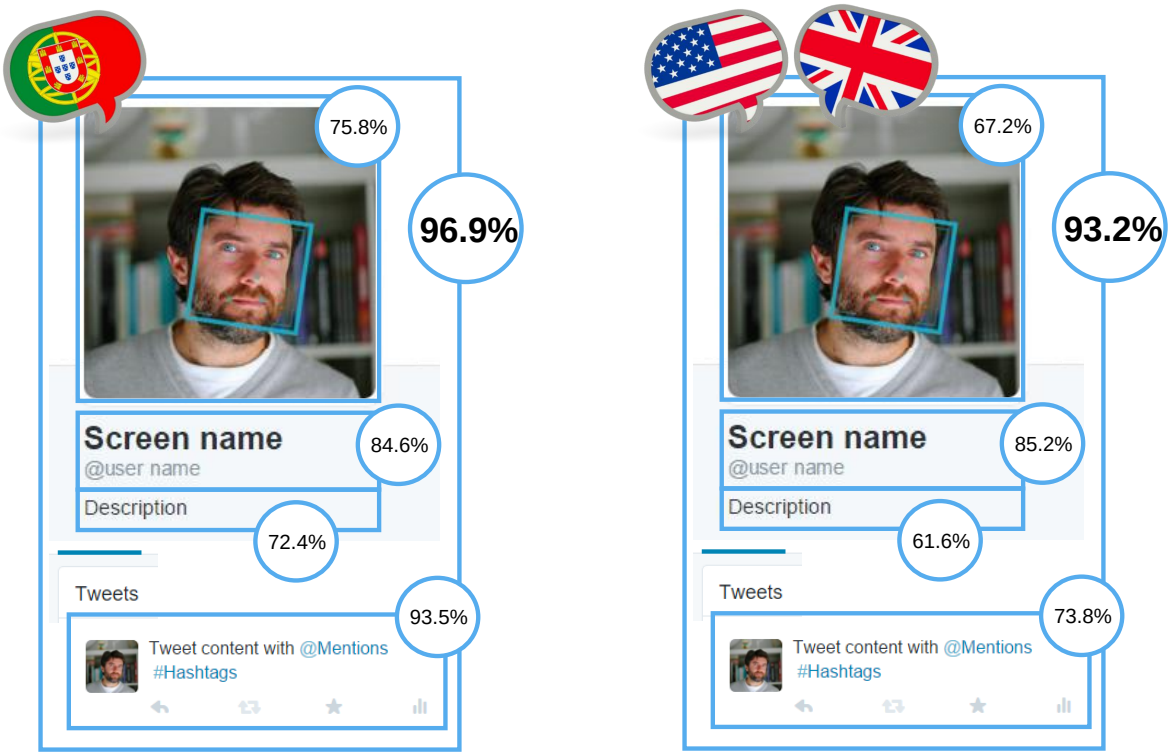


Figure 6.5: Classification accuracy per group of features for both datasets.

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. Figure 6.5 summarizes the achieved accuracies per classifier for both datasets.

7 Conclusions and future work

Would you tell me, please, which way I ought to go from here?

That depends a good deal on where you want to get to, said the Cat.

I don't much care where - said Alice.

Then it doesn't matter which way you go, said the Cat.

Lewis Carroll, *Alice in Wonderland*

This chapter overviews the work reported in this thesis, presents the main conclusions, enumerates the main contributions, and describes a number of possible directions for further extending this research.

7.1 Conclusions

In this study, we began by reviewing the existing approaches to overcome the problem of Twitter gender classification. In previous research, the most common features are based on textual content, profile information and social networking. Textual content, including the user bio, is usually classified using ngrams, both from characters and words. Character ngrams are very useful when analyzing languages like Japanese, where a word can be represented by a single character. Profile information is used basically by searching for names related to a specific gender, using dictionaries of names. Social network features tend to be weak. Al Zamal et al. (2012) 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. The profile picture is usually disregarded, though it might provides clues to users' gender. One common problem in previous works, is the inexistence of labelled corpora. This means imposing a labor intensive task of manually labelling users. Consequently, some studies

use small labelled datasets for the creation of their models. Usually the task of labelling is performed by looking for names in the profile information. Most of the previous work has been done for the English language. To our best knowledge there is no study applied to Portuguese users.

After reviewing previous research, we experimented a method to automatically detect user's gender uniquely based on unstructured information available in the user's profile. We started by extracting data from Twitter and created a dataset of English users and a dataset of Portuguese users. After the creation of the dataset, we manually labelled about 700 users for each dataset. We compiled two dictionaries of names with the corresponding gender. Unisex names were ignored. With the aid of the dictionaries, we extracted features associating names found in the *user name* and *screen name* with the corresponding gender. We evaluated the performance of the features using several supervised and unsupervised approaches, including Naive Bayes variants, Logistic Regression, Support Vector Machines, Fuzzy c-Means clustering, and k-means. Results show that features perform well in both languages. Supervised approaches reached 97.9% accuracy, but Fuzzy c-Means also proved suitable for this task achieving 96.4% accuracy. We noticed that using unsupervised methods, the increasing amount of data has positive impact on the results. The only restriction for this method is that within the user profile there is at least a sequence of characters matching a name contained within a dictionary. For those users, we obtained an accuracy of 97.9% for English users and 98.3% for Portuguese users.

Our next step was to create extended labelled datasets in a semi-automatic fashion, based on our proposed profile features. We created an English and a Portuguese dataset, filtering users by tweets' language, each composed of about 100k users. For each user, we extracted the last 100 tweets. After the creation of the datasets, we classified users with the Multinomial Naive Bayes model from our previous experiments. All users classified with a confidence superior or equal to 95% were kept. The resulting Portuguese labelled dataset had 58k labelled users, while the English dataset had 65k labelled users. It is important to notice that in our first experiments, we only had about 700 manually labelled users in each dataset. In order to enhance the datasets further, we added two new attributes for each user: gender recognition from profile picture, and detailed geographical information based on the last known location. The first attribute provides useful information for the gender classification, while the second attribute is relevant for tackling region specific phenomena. Finally, we manually validated a random sample of the users' classification to insure the quality of our dataset. In the English dataset we detected 9% of incorrect gender classification, while in the Portuguese dataset only 3% of the users were classified incorrectly. These results show the success of the proposed method for labelled datasets creation.

Finally, 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 the extended labelled datasets from our previous experiments, 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. MNB 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, MNB 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. SVM 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 goals: we created a semi-automatic gender labelled dataset, 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.

7.2 Future work

Even though the combined classifier achieves high accuracy in both languages, our goal is to improve the combined classifier, adding new features. Future work will also encompass the classification of other latent user attributes. Namely, age, football club preference and political affiliation. Using our Twitter gender labelled dataset, we will also investigate the possible relation between age and gender in the language usage in Twitter.

Bibliography

Al Zamal, F., Liu, W., and Ruths, D. (2012). Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. *ICWSM*, 270.

Alowibdi, J. S., Buy, U., Yu, P., et al. (2013). Empirical evaluation of profile characteristics for gender classification on twitter. In *Machine Learning and Applications (ICMLA), 2013 12th International Conference on*, volume 1, pages 365–369. IEEE.

Alowibdi, J. S., Buy, U. A., and Philip, S. Y. (2014). Say it with colors: Language-independent gender classification on twitter. In *Online Social Media Analysis and Visualization*, pages 47–62. Springer.

Aravantinou, C., Simaki, V., Mporas, I., and Megalooikonomou, V. (2015). Gender classification of web authors using feature selection and language models. In Ronzhin, A., Potapova, R., and Fakotakis, N., editors, *Speech and Computer*, volume 9319 of *Lecture Notes in Computer Science*, pages 226–233. Springer International Publishing.

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

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

Arroju, M., Hassan, A., and Farnadi, G. (2015). Age, gender and personality recognition using tweets in a multilingual setting.

Bamman, D., Eisenstein, J., and Schnoebelen, T. (2012). Gender in twitter: Styles, stances, and social networks. *CoRR abs/1210.4567*.

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

Baptista, J., Batista, F., Mamede, N. J., and Mota, C. (2005). Npro: um novo recurso para o processamento computacional do português. In *XXI Encontro APL*.

Baumann, A., Krasnova, H., Veltri, N. F., and Ye, Y. (2015). Men, women, microblogging: Where do we stand?

Bechar-Israeli, H. (1995). From< bonehead> to< clonehead>: Nicknames, play, and identity on internet relay chat1. *Journal of Computer-Mediated Communication*, 1(2):0–0.

Bei, J. (2013). How chinese journalists use weibo microblogging for investigative reporting.

Bergsma, S., Dredze, M., Van Durme, B., Wilson, T., and Yarowsky, D. (2013). Broadly improving user classification via communication-based name and location clustering on twitter. In *HLT-NAACL*, pages 1010–1019.

Bezdek, J. C., Ehrlich, R., and Full, W. (1984). Fcm: The fuzzy c-means clustering algorithm. *Computers and Geosciences*, 10(2–3):191 – 203.

Brogueira, G., Batista, F., Carvalho, J. P., and Moniz, H. (2014). Expanding a database of portuguese tweets. In Pereira, M. J. V., Leal, J. P., and Simões, A., editors, *3rd Symposium on Languages, Applications and Technologies*, volume 38 of *OpenAccess Series in Informatics (OASIcs)*, pages 275–282, Dagstuhl, Germany. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

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

Burger, J. D., Henderson, J., Kim, G., and Zarrella, G. (2011). Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1301–1309, Stroudsburg, PA, USA. Association for Computational Linguistics.

Callison-Burch, C. and Dredze, M. (2010). Creating speech and language data with amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 1–12. Association for Computational Linguistics.

Calvert, S. L., Mahler, B. A., Zehnder, S. M., Jenkins, A., and Lee, M. S. (2003). 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.

Carvalho, J. P., Pedro, V., and Batista, F. (2013). Towards intelligent mining of public social networks' influence in society. In *IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS)*, pages 478 – 483, Edmonton, Canada.

Chen, Y., You, J., Chu, M., Zhao, Y., and Wang, J. (2006). Identifying language origin of person names with n-grams of different units. In *IEEE ICASSP 2006.*, volume 1, pages I–I.

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

Cieri, C., Miller, D., and Walker, K. (2004). The fisher corpus: a resource for the next generations of speech-to-text. In *LREC*, volume 4, pages 69–71.

Ciot, M., Sonderegger, M., and Ruths, D. (2013). Gender inference of twitter users in non-english contexts. In *EMNLP*, pages 1136–1145.

Conover, M., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F., and Flammini, A. (2011). Political polarization on twitter. In *ICWSM*.

Consortium, O. et al. (1995). The onomastica interlanguage pronunciation lexicon.

Corney, M. W. (2003). *Analysing e-mail text authorship for forensic purposes*. PhD thesis, Queensland University of Technology.

Culotta, A. (2010). Detecting influenza outbreaks by analyzing twitter messages. *arXiv preprint arXiv:1007.4748*.

Deitrick, W., Miller, Z., Valyou, B., Dickinson, B., Munson, T., and Hu, W. (2012). Gender identification on twitter using the modified balanced winnow.

Earle, P., Guy, M., Buckmaster, R., Ostrum, C., Horvath, S., and Vaughan, A. (2010). Omg earthquake! can twitter improve earthquake response? *Seismological Research Letters*, 81(2):246–251.

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

Eisenstein, J., O'Connor, B., Smith, N. A., and Xing, E. P. (2010). A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287. Association for Computational Linguistics.

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

Finger, L. (2015). Do evil - the business of social media bots. http://www.forbes.com/sites/lutzfinger/2015/02/17/do-evil-the-business-of-social-media-bots/. (Visited on 21/02/2015).

Fink, C., Kopecky, J., and Morawski, M. (2012). Inferring gender from the content of tweets: A region specific example. In *ICWSM*.

Fischer, J. L. (1958). Social influences on the choice of a linguistic variant. *Word*, 14(1):47–56.

Garera, N. and Yarowsky, D. (2009). Modeling latent biographic attributes in conversational genres. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 710–718, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Goswami, S. and Shishodia, M. (2012). A fuzzy based approach to stylometric analysis of blogger's age and gender. In *Hybrid Intelligent Systems (HIS), 2012 12th International Conference on*, pages 47–51.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18.

Halteren, H. v. and Speerstra, N. (2014). Gender recognition on dutch tweets.

Hanneman, R. A. and Riddle, M. (2005). Introduction to social network methods.

Harris, Z. S. (1954). Distributional structure. *Word*.

Heil, B. and Piskorski, M. (2009). New twitter research: Men follow men and nobody tweets. *Harvard Business Review*, 1:2009.

Holmes, J. and Meyerhoff, M. (2008). *The handbook of language and gender*, volume 25. John Wiley & Sons.

Huffaker, D. (2004). *Gender similarities and differences in online identity and language use among teenage bloggers*. PhD thesis, Citeseer.

Imran, M., Castillo, C., Diaz, F., and Vieweg, S. (2015). Processing social media messages in mass emergency: A survey. *ACM Computing Surveys (CSUR)*, 47(4):67.

Jaech, A. and Ostendorf, M. (2015). What your username says about you. *arXiv preprint arXiv:1507.02045*.

Jamali, M. and Abolhassani, H. (2006). Different aspects of social network analysis. In *Web Intelligence, 2006. WI 2006. IEEE/WIC/ACM International Conference on*, pages 66–72. IEEE.

Joachims, T. (1999). Making large scale svm learning practical. Technical report, Universität Dortmund.

Kaisser, M. and Lowe, J. (2008). Creating a research collection of question answer sentence pairs with amazon's mechanical turk. In *LREC*.

Keerthi, S. S., Shevade, S. K., Bhattacharyya, C., and Murthy, K. R. K. (2001). Improvements to platt's smo algorithm for svm classifier design. *Neural Computation*, 13(3):637–649.

Kokkos, A. and Tzouramanis, T. (2014). A robust gender inference model for online social networks and its application to linkedin and twitter. *First Monday*, 19(9).

Koppel, M., Argamon, S., and Shimoni, A. R. (2002). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412.

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

Labov, W. (2006). *The social stratification of English in New York city*. Cambridge University Press.

Le Cessie, S. and Van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Applied statistics*, pages 191–201.

Leetaru, K., Wang, S., Cao, G., Padmanabhan, A., and Shook, E. (2013). Mapping the global twitter heartbeat: The geography of twitter. *First Monday*, 18(5).

Liu, W., Al Zamal, F., and Ruths, D. (2012). 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*.

Liu, W. and Ruths, D. (2013). What's in a name? using first names as features for gender inference in twitter. In *AAAI Spring Symposium: Analyzing Microtext*.

Lotan, G., Graeff, E., Ananny, M., Gaffney, D., Pearce, I., et al. (2011). The arab spring| the revolutions were tweeted: Information flows during the 2011 tunisian and egyptian revolutions. *International journal of communication*, 5:31.

Ludu, P. S. (2014). Inferring gender of a twitter user using celebrities it follows. *arXiv preprint arXiv:1405.6667*.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations.

McCallum, A., Nigam, K., et al. (1998). A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Citeseer.

McCreadie, R., Soboroff, I., Lin, J., Macdonald, C., Ounis, I., and McCullough, D. (2012). On building a reusable twitter corpus. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 1113–1114. ACM.

McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444.

Merler, M., Cao, L., and Smith, J. R. (2015). 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*, pages 1–6. IEEE.

Miller, Z., Dickinson, B., and Hu, W. (2012). Gender prediction on twitter using stream algorithms with n-gram character features.

Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., and Rosenquist, J. N. (2011). Understanding the demographics of twitter users. *ICWSM*, 11:5th.

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

Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, 45(2):167–256.

Nguyen, D., Trieschnigg, D., Dogruöz, A. S., Gravel, R., Theune, M., Meder, T., and de Jong, F. (2014). Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment.

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

Pennacchiotti, M. and Popescu, A.-M. (2011). A machine learning approach to twitter user classification. *ICWSM*, 11:281–288.

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., and Booth, R. J. (2007). The development and psychometric properties of liwc2007.

Petrović, S., Osborne, M., and Lavrenko, V. (2010). The edinburgh twitter corpus. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, WSA '10, pages 25–26, Stroudsburg, PA, USA. Association for Computational Linguistics.

Platt, J. et al. (1999). Fast training of support vector machines using sequential minimal optimization. *Advances in kernel methods—support vector learning*, 3.

Rao, D., Yarowsky, D., Shreevats, A., and Gupta, M. (2010). Classifying latent user attributes in twitter. In *Proceedings of the 2Nd International Workshop on Search and Mining User-generated Contents*, SMUC '10, pages 37–44, New York, NY, USA. ACM.

Rocha, P. and Santos, D. (2000). Cetempúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. *Actas do V Encontro para o processamento computacional da língua portuguesa escrita e falada, PROPOR*, 2000:131–140.

Rosa, H., Batista, F., and Carvalho, J. P. (2014). Twitter topic fuzzy fingerprints. In *WCCI2014, FUZZ-IEEE, 2014 IEEE World Congress on Computational Intelligence, International Conference on Fuzzy Systems*, IEEE Xplorer, pages 776–783, Beijing, China.

Schler, J., Koppel, M., Argamon, S., and Pennebaker, J. W. (2006). Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, volume 6, pages 199–205.

Stopczynski, A., Sekara, V., Sapiezynski, P., Cuttone, A., Madsen, M. M., Larsen, J. E., and Lehmann, S. (2014). Measuring large-scale social networks with high resolution. *PloS one*, 9(4):e95978.

Ugheoke, T. O. (2014). Detecting the gender of a tweet sender.

Van Zegbroeck, E. (2014). Predicting the gender of flemish twitter users using an ensemble of classifiers.

Vieweg, S., Hughes, A. L., Starbird, K., and Palen, L. (2010). 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*, pages 1079–1088. ACM.

Wauters, R. (2010). Only 50% of twitter messages are in english, study says. *TechCrunch. com, http://techcrunch. com/2010/02/24/twitter-languages*.

You, Q., Bhatia, S., Sun, T., and Luo, J. (2014). The eyes of the beholder: Gender prediction using images posted in online social networks. In *Data Mining Workshop (ICDMW), 2014 IEEE International Conference on*, pages 1026–1030. IEEE.