Twitter gender classification using user unstructured information

Marco Vicente
INESC-ID,
ISCTE-IUL, Lisboa, Portugal
Email: mpfve@iscte.pt

Fernando Batista
INESC-ID,
ISCTE-IUL, Lisboa, Portugal
Email: fernando.batista@iscte.pt

Joao Paulo Carvalho
INESC-ID, Instituto Superior Técnico
Universidade de Lisboa, Portugal
Email: joao.carvalho@inesc-id.pt

***Abstract*—This paper describes an approach to automatically detect the gender of Twitter users, based only on clues provided by their profile information in an unstructured form. A number of features that capture phenomena specific of Twitter users is proposed and evaluated on a dataset of about 242K English language users. Different supervised and unsupervised approaches are used to assess the performance of the proposed features, including Naive Bayes variants, Logistic Regression, Support Vector Machines, Fuzzy c-Means clustering, and K-means. An unsupervised approach based on Fuzzy c-Means proved to be very suitable for this task, returning the correct gender for about 96% of the users.**

***Index Terms*—Twitter; Gender detection; Fuzzy c-Means; Supervised and unsupervised methods.**

I. Introduction

With the massification of social networks, social media has become a playground for researchers. Among public social networks, Twitter, with 288 million monthly active users and 500 million tweets sent per day [1], has become a major tool for social networking studies [2], [3]. Researchers are mining content generated in Twitter to understand public opinion (sentiment analysis, political activity), to monitor diseases (e.g. detect flu outbreaks [4]) or even to improve response to natural catastrophes (e.g. detect earthquakes [5]).

The information provided by Twitter about a user is limited and does not specifically includes relevant information, such as gender or age. 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 the users community [6]. In fact, age and gender information is most of the times provided wittingly or unwittingly by the user, but it is available in an unstructured form. Unlike other social networks, when creating a Twitter profile, the only required field when creating an account is a user name. Additionally, a user profile includes the following optional attributes that can be changed by the user without restrictions [7], and that may provide additional information about the user in an unstructured form:

- Screen name (e.g.: johndoe95)
- User name (e.g.: John Doe the best :))
- Location
- URL
- Description

This paper proposes a method to automatically detect the user's gender (male or female), 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 are proposed for extracting useful information from the user's profile attributes, namely *user name,* and *screen name*. The online content generated by the user in each tweet is not used in the scope of this work. 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 that users tend to choose real names more often than other forms [8], [9]. Several methods were applied to the extracted features in order to automatically obtain user gender. It is shown that by using Fuzzy C-means [10] it is possible to obtain state of the art results recurring to an unsupervised learning method that excludes the need to label a training set.

This paper is organized as follows: Section II overviews previous work concerning gender detection. Section III describes the proposed feature extraction method. Section IV characterizes the data and describes the process of manually labelling the gender of a subset of the users. Section V describes our experiments using supervised and unsupervised methods and reports the corresponding results. Section VI presents the conclusions and prospects about the future work.

II. Related Work

The problem of deciding whether the Twitter user is male or female, simply based on the Twitter user profile content has been rarely addressed in the literature. However, a related 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 [11], [12], [13], [14].

The study of the relation between gender and language usage is extensive (for an overview, see e.g.: Holmes et al [15] and Eckert et al [16]). Research has been published which supports the hypothesis that by analyzing linguistic features associated with male or female speech, it is possible to detect users' gender by their use of language [17]. Kopel et al [11],

using automated text categorization techniques, report gender detection with approximately 80% accuracy using function words and parts of speech.

In a later research [18], 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 used by Koppel et al [12] to gender detection. They report an overall accuracy of 76.1% using word classes derived from systemic functional linguistics and character ngrams. This corpus was used by Goswami et al [13]. They improved the overall accuracy to 89.2%, using average sentence length, usage of slang and usage of non-dictionary words.

Gender detection has been applied to Twitter. Rao et al [19] examined Tweets written in English, using Support Vector Machines with character ngram-features and sociolinguistic features like emoticons use or alphabetic character repetitions. They reported an accuracy of 71,8% using sociolinguistic features, using ngrams they reached only an accuracy of 67.7%. When combining ngram-features with sociolinguistic features, the accuracy reached 72.3%.

The state-of-the-art study of Burger et al [20] collects a large multilingual dataset labelled with gender. While Rao et al[19] used only a manual annotation of 500 English users labelled with gender, Burger et al created a corpus of approximately 184000 Twitter users labelled with gender, with a training set of 3.3 million tweets and a test set of 418000 tweets. They used Support Vector Machines, Naive Bayes and Balanced Winnow2 with word and character N-grams as features to detect gender. Using tweet texts alone they achieved the accuracy of 75.5%. When combining tweet texts with profile information (description, user name and screen name), they achieved 92.0% of accuracy.

Fink et al [21] studied tweets from Nigeria using Support Vector Machine with a linear kernel implementation [22]. Using unigram features (word unigrams, hash tags, and psychometric properties) from tweet texts alone, they obtained an accuracy of 80.5% predicting gender. Bamman et al [23] 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). Halteren et al [24]studied a corpus of Dutch tweets of 600 labelled users. Using Tweet text only (using both character and token n-grams), they achieved an accuracy of 95.5% detecting gender. The machine learning system used was Support Vector Regression with a 5-fold cross-validation on the corpus. In their research Linguistic Profiling and TiMBL are used, but with inferior accuracy.

Other studies on gender detection report results of stylometric differences in blogs for gender and age group variation [13]. [14] presents a follow-up of the previous work using Fuzzy c-means to detect age and gender on blog's text, achieving a peak accuracy of around 84% for gender estimation.

III. Proposed Method

This section describes the process used to estimate the gender of a Twitter user, given only the unstructured information available for each tweet in the user's profile. Our approach consists of expressing the profile information by a set of features that are then used to estimate the gender of the user, in both supervised and unsupervised fashions. The features are extracted with the help of a dictionary of names containing the corresponding gender and a given frequency. The following subsections describe in detail the process of compiling the dictionary and the feature extraction process.

A. Names Dictionary

In order to automatically associate names that can be possibly found in the user's profile with the corresponding gender, we have compiled a dictionary of about 8444 names. Such dictionary was constructed based on the list of the most used baby names from the United States Social Security Administration Official Website [25]. 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 are 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 paper, we extracted only data from 1940 and beyond, where the name occurred at least 1000 times. We focused on names that are exclusively male or female, since unisex names can be classified as male or female. We created a dictionary with the following information:

- name
- gender
- number of occurrences

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.

B. 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, but it usually does not contain names of the user, making it less suitable for this task. Our suggested gender feature extraction algorithm is

Algorithm 1 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

summarily described in Algorithm 1 and illustrated in Figure 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.
  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 [26]. 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 I).

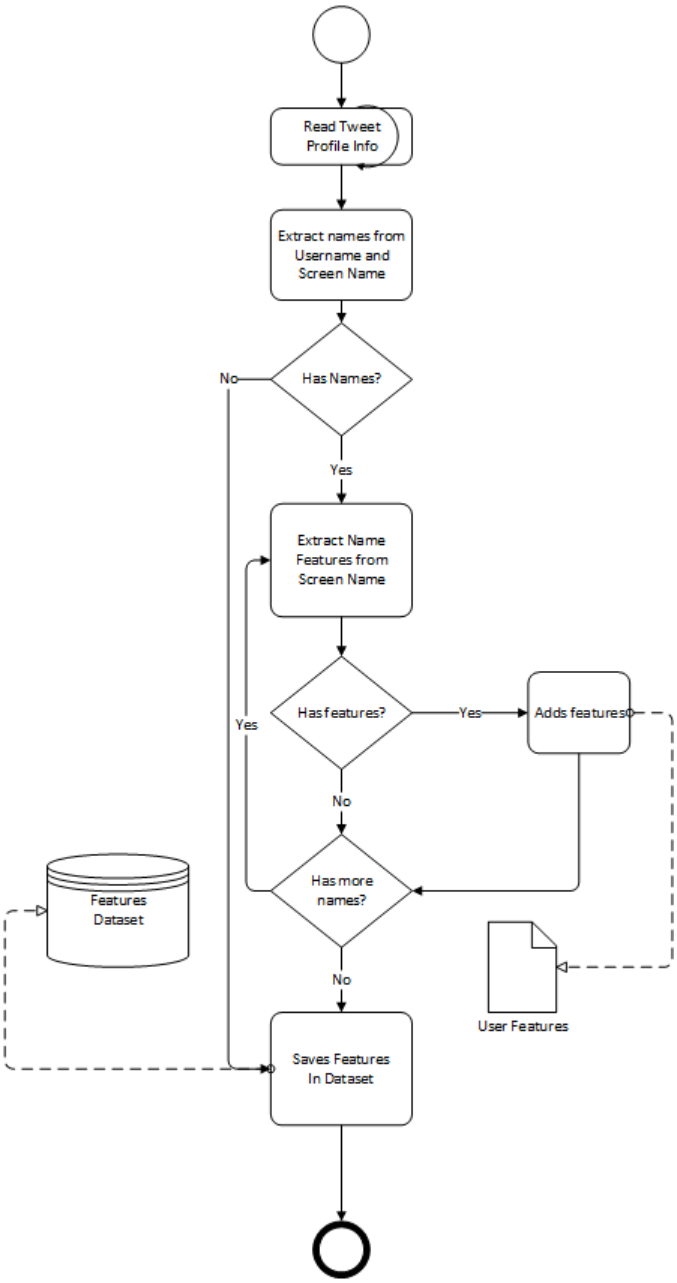

Figure 1. Feature extraction diagram

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").

In the research of Burger et al [20], n-grams of profile

Table I
LEET SPEAK REPLACEMENTS

LEET	CHARACTER
3	e
1	l
0	o
7	t
4	a
6	g
$	s

Table II
A SELECTION OF EXISTING FEATURES

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

user name are used to infer gender. We suggest a different approach: 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"
- **Boundaries**: indicates if individual words in screen name are properly bounded, i.e., if they start with a capital, 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 II 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"*.

Two different modes were considered for feature extraction: Lazy and Greedy. Lazy mode extracts only the more granular feature for each name. Greedy mode extracts all applicable features for each name. Consider the name "Ernest Hemingway" as an example:

greedy 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"

lazy feature extraction:

"name_beginning_with_separation_and_case_m"

Table III
DATASET OF ENGLISH TWITTER USERS.

	U.S.	U.K.	Total
Number of users	199950	42708	242658
Screen Name average length	5.36	5.24	5.34
User Name average length	5.41	5.2	5.37
User Name avg number of words	1.31	1.37	1.32

IV. DATASET

The dataset used in this paper was extracted from one month of tweets collected during December of 2014, using the Twitter *streaming/sample* API [7], [27]. This method gives access to only about 1% of the actual public tweets [28]. We have restricted the data to English and geolocated tweets, 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. For the purpose of this paper, the dataset has been further restricted to users for whom their profile information matched at least one of the gender features previously mentioned in Subsection III-B (82% of the users). Table III shows the resultant dataset, distributed by country, which has been used in our research.

A. Labelled data

In order to evaluate our method of gender classification, we manually labelled a randomly selected portion of the users. Each one of the users was associated with the corresponding gender.

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) [19], [29], [30]. In the research

Table IV
LABELLED DATA PROPERTIES BY GENDER.

	Male	**Female**	**Total**
#Users	330	418	748
#Leet occurrences	8	51	59
#Repeated vowels	2	8	10
Number of extracted features	2168	2978	5146

of Ciot et al [31], the gender is identified using the profile picture associated with the user account. The study of Burger et al [20] complementarily examines blog sites (found in the URL field in their profile) to label users with gender. In fact, the research of Huffaker [32] 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 subset 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.

1) Firstly, we looked for names both in the *user name* and in the *screen name* of the profile. It is worth noting that, as previously mentioned, all the users in our database contain at least a sequence that matches a name in our dictionary of names. If our manual evaluation of the extracted names turned out to be 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, based on previous findings that state that about 7% of tweet profiles are non-human spam bots [33]. 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.

By applying this method we created a subset with 748 manually labelled users for whom we were able to determine their gender. Table IV characterizes the labelled data subset, revealing a difference between male (44%) and female (56%) users. The data is consistent with a previous study of correlation between name and gender that estimates Twitter has about 45% of male users [34]. Notwithstanding, there is a predominance of "leet speak" and repeated vowels usage in female users (86% and 80% respectively).

Table V shows the number of features that can be extracted from the manually labelled subset. 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.

Table V
FEATURES EXTRACTED FROM EACH PROFILE ATTRIBUTE.

	User name	Screen Name
Number of extracted features	3221	1925
Leet related features	291	208
Repeated vowels related features	20	48

Table VI
PROPERTIES OF THE EXTRACTED NAMES.

	User Name	Screen Name
Average Name Length (chars)	5.4	5.3
Percentage of rejected names	29%	73%

Table VI shows statistics for the extracted names in each one of the profile attributes, revealing 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.

V. GENDER CLASSIFICATION EXPERIMENTS

This section describes our experiments using the proposed features and different supervised and unsupervised approaches, which make it possible to assess the performance of the proposed features. The supervised methods include: Naive Bayes variants, Logistic Regression, and Support Vector Machines. The unsupervised methods include Fuzzy c-Means clustering [10] and k-means [35]. The Fuzzy c-Means method uses the fuzzy logic toolkit for SciPy [36] that can be found at https://github.com/scikit-fuzzy/scikit-fuzzy. All the other methods were applied through Weka[1], a collection of open source machine learning algorithms and a collection of tools for data pre-processing and visualization [37].

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. Two different ways of extracting the features are being used (as explained in Section III-B):

1) lazy feature extraction - uses only the most granular feature for each name;
2) greedy feature extraction - uses all possible triggered features.

All experiments use the extracted features in a binary fashion.

A. Supervised classification

Experiments using supervised methods use the labelled data for training and evaluating the models using a 5-fold cross-validation. Different methods were compared to assess the

[1]Weka version 3-6-8. http://www.cs.waikato.ac.nz/ml/weka

Table VII
CLASSIFICATION RESULTS FOR SUPERVISED METHODS.

	Lazy features		Greedy features	
	Accuracy	kappa	Accuracy	kappa
Logistic Regression	94.5%	0.89	93.7 %	0.87
Multinomial Naive Bayes	**97.2%**	**0.94**	**97.2%**	**0.94**
Support Vector Machines	96.0%	0.92	96.4%	0.93

Table VIII
CLASSIFICATION ACCURACY FOR UNSUPERVISED METHODS.

	Lazy features		Greedy features	
	labelled	all data	labelled	all data
kMeans clustering	74.9%	71.2%	75.1%	67.3%
Fuzzy c-Means	84.9%	87.3%	93.1%	**96.0%**

performance of the proposed features, namely: Multinomial Naive Bayes (MNB) [38], a variant of Naive Bayes, Logistic Regression [39], and Support Vector Machines [40], [41].

Table VII summarizes the results achieved with each one of the methods in the task of distinguishing between male and female users. The Multinomial Naive Bayes method achieved the best performance using our features. Support Vector Machines are still better than Logistic Regression for this task, but are still about 1% lower than MNB. The kappa statistic is a chance-corrected measure of agreement between the classifications and the true classes. A value close to zero indicates that results could be achieved almost by chance whereas a value close to 1 means an almost complete agreement, and reveals a suitable model for the problem.

Our proposed features prove to be good for discriminating the user's gender in Twitter, achieving a performance of about 97% accuracy when using a supervised approach.

B. Unsupervised classification

Fuzzy c-Means clustering and K-means clustering were the two unsupervised methods applied. k-means was set to use the Euclidean distance, the centroids are computed as a mean, the number of clusters has been set to 2, and the seed was set to 10 (default value). In order to use the Fuzzy c-means clustering algorithm, the data has been converted into a matrix of binary values, and parameters were as follows:

- Number of clusters: 2
- Maximum number of iterations: 1000
- Distance function: Euclidean

Two different experiments are being performed in order to assess the impact of the amount of data in the performance: the first experiment uses only the dataset of labelled data for building the clusters and the same data for evaluating the classification performance; the second experiment uses all unlabelled data for clustering and uses labelled for evaluating the results. Table VIII summarizes the results, and reveals that Fuzzy c-Means proved to be very suitable for this task, returning the correct gender for about 96% of the users when all the data is used for learning the clusters. The performance of Fuzzy c-Means significantly increased when more data was used for learning the clusters and also when more, but less robust, features were added. These results suggest that even better performance could be achieved by using more unlabelled data. On the other hand, the performance using kMeans is always lower than Fuzzy c-Means, and gets even worse when more unlabelled data is provided.

Fuzzy c-means proved to be an excellent choice for the gender detection on Twitter since:

1) it does not require labelled data, something that is fundamental when dealing with Twitter
2) its performance increases as more data is provided
3) it achieves a performance almost similar (1% absolute) to the best supervised method.

Despite not being directly comparable because datasets are different, our proposed features compare well with the performance achieved by other state-of-the art research. For example, [20] uses the winnow algorithm with n-grams extracted from the user's full name and obtain 89.1% accuracy for gender detection. The proposed features can be used to extend gender labelled datasets for researchers.

VI. CONCLUSIONS AND FUTURE WORK

We have described an approach to automatically detect the gender of Twitter users, based on clues provided by their profile information. A number of name related features that capture phenomena specific of Twitter users is proposed and evaluated on a dataset of about 242K English users. Different supervised and unsupervised approaches are used to assess the performance of the proposed features, including Naive Bayes variants, Logistic Regression, Support Vector Machines, Fuzzy c-Means clustering, and K-means. Our proposed features proved to be good for discriminating the user's gender in Twitter, achieving a performance of about 97% accuracy when using a supervised approach. The unsupervised approach based on Fuzzy c-Means proved also to be very suitable for this task, returning the correct gender for about 96% of the users, with the added advantages of not needing a labelled training set and of possible accuracy improvements with larger datasets. It should be noted that reported results assume that at least one usable feature was triggered from the unstructured information (82% of all users).

Future work will encompass the creation of extended labelled datasets in a semi-automatic fashion, based on an automatic annotation provided by our proposed features. Such extended labelled datasets will make it possible to associate the textual content provided by the users with their gender and create gender models, purely based on the text contents. Such models, based on huge amounts of data, can then be adapted and used in a cross-domain scenario.

ACKNOWLEDGEMENTS

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

REFERENCES

[1] Twitter, "Twitter usage." https://about.twitter.com/company.

[2] G. Brogueira, F. Batista, J. P. Carvalho, and H. Moniz, "Expanding a Database of Portuguese Tweets," in *3rd Symposium on Languages, Applications and Technologies* (M. J. V. Pereira, J. P. Leal, and A. Simões, eds.), vol. 38 of *OpenAccess Series in Informatics (OASIcs)*, (Dagstuhl, Germany), pp. 275–282, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2014.

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

[4] A. Culotta, "Detecting influenza outbreaks by analyzing twitter messages," *arXiv preprint arXiv:1007.4748*, 2010.

[5] P. Earle, M. Guy, R. Buckmaster, C. Ostrum, S. Horvath, and A. Vaughan, "Omg earthquake! can twitter improve earthquake response?," *Seismological Research Letters*, vol. 81, no. 2, pp. 246–251, 2010.

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

[7] T. development team, "Api overview." https://dev.twitter.com/overview/api, 2015.

[8] H. Bechar-Israeli, "From< bonehead> to< clonehead>: Nicknames, play, and identity on internet relay chat1," *Journal of Computer-Mediated Communication*, vol. 1, no. 2, pp. 0–0, 1995.

[9] S. L. Calvert, B. A. Mahler, S. M. Zehnder, A. Jenkins, and M. S. Lee, "Gender differences in preadolescent children's online interactions: Symbolic modes of self-presentation and self-expression," *Journal of Applied Developmental Psychology*, vol. 24, no. 6, pp. 627–644, 2003.

[10] J. C. Bezdek, R. Ehrlich, and W. Full, "Fcm: The fuzzy c-means clustering algorithm," *Computers and Geosciences*, vol. 10, no. 2–3, pp. 191 – 203, 1984.

[11] M. Koppel, S. Argamon, and A. R. Shimoni, "Automatically categorizing written texts by author gender," *Literary and Linguistic Computing*, vol. 17, no. 4, pp. 401–412, 2002.

[12] M. Koppel, J. Schler, and S. Argamon, "Computational methods in authorship attribution," *Journal of the American Society for information Science and Technology*, vol. 60, no. 1, pp. 9–26, 2009.

[13] S. Goswami, S. Sarkar, and M. Rustagi, "Stylometric analysis of bloggers age and gender," in *Third International AAAI Conference on Weblogs and Social Media*, 2009.

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

[15] J. Holmes and M. Meyerhoff, *The handbook of language and gender*, vol. 25. John Wiley & Sons, 2008.

[16] P. Eckert and S. McConnell-Ginet, *Language and gender*. Cambridge University Press, 2013.

[17] M. Bucholtz and K. Hall, "Identity and interaction: A sociocultural linguistic approach," *Discourse studies*, vol. 7, no. 4-5, pp. 585–614, 2005.

[18] J. Schler, M. Koppel, S. Argamon, and J. W. Pennebaker, "Effects of age and gender on blogging.," in *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, vol. 6, pp. 199–205, 2006.

[19] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta, "Classifying latent user attributes in twitter," in *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pp. 37–44, ACM, 2010.

[20] J. D. Burger, J. Henderson, G. Kim, and G. Zarrella, "Discriminating gender on twitter," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1301–1309, Association for Computational Linguistics, 2011.

[21] C. Fink, J. Kopecky, and M. Morawski, "Inferring gender from the content of tweets: A region specific example.," in *ICWSM*, 2012.

[22] T. Joachims, "Making large scale svm learning practical," tech. rep., Universität Dortmund, 1999.

[23] D. Bamman, J. Eisenstein, and T. Schnoebelen, "Gender identity and lexical variation in social media," *Journal of Sociolinguistics*, vol. 18, no. 2, pp. 135–160, 2014.

[24] H. v. Halteren and N. Speerstra, "Gender recognition on dutch tweets," 2014.

[25] U. S. S. S. Administration, "Popular baby names." http://www.ssa.gov/oact/babynames/limits.html, 2015.

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

[27] T. development team, "Twitter dev tools." https://dev.twitter.com/, 2015.

[28] T. development team, "Limit on streaming tweets." https://dev.twitter.com/discussions/6789, 2015.

[29] M. Pennacchiotti and A.-M. Popescu, "A machine learning approach to twitter user classification.," *ICWSM*, vol. 11, pp. 281–288, 2011.

[30] F. Al Zamal, W. Liu, and D. Ruths, "Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors.," *ICWSM*, vol. 270, 2012.

[31] M. Ciot, M. Sonderegger, and D. Ruths, "Gender inference of twitter users in non-english contexts.," in *EMNLP*, pp. 1136–1145, 2013.

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

[33] L. Finger, "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/, 2015. (Visited on 02/21/2015).

[34] B. Heil and M. Piskorski, "New twitter research: Men follow men and nobody tweets," *Harvard Business Review*, vol. 1, p. 2009, 2009.

[35] J. MacQueen, "Some methods for classification and analysis of multivariate observations," 1967.

[36] J. Warner, "Scikit fuzzy - fuzzy logic toolbox for python." https://github.com/scikit-fuzzy/scikit-fuzzy.

[37] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explor. Newsl.*, vol. 11, pp. 10–18, Nov. 2009.

[38] A. McCallum, K. Nigam, *et al.*, "A comparison of event models for naive bayes text classification," in *AAAI-98 workshop on learning for text categorization*, vol. 752, pp. 41–48, Citeseer, 1998.

[39] S. Le Cessie and J. C. Van Houwelingen, "Ridge estimators in logistic regression," *Applied statistics*, pp. 191–201, 1992.

[40] J. Platt *et al.*, "Fast training of support vector machines using sequential minimal optimization," *Advances in kernel methods—support vector learning*, vol. 3, 1999.

[41] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy, "Improvements to platt's smo algorithm for svm classifier design," *Neural Computation*, vol. 13, no. 3, pp. 637–649, 2001.