WHERE ARE YOU FROM?
Tell me HOW you Write and I Will Tell you WHO you Are

Marta R. Costa-jussà, Rafael E. Banchs and Joan Codina
Barcelona Media Research Center, Av Diagonal 177, 9th floor, 08018 Barcelona, Spain

Keywords: Data mining, Bag-of-words classification.

Abstract: The main goal of this study is evaluating the feasibility of predicting the country, the age and the gender of a social network user, given his/her posts in one or several forums. The study was conducted with MySpace forums in Spanish language, aiming at classifying and extracting demographic information along with age and gender differences. The preliminary results presented and discussed here show some interesting conclusions about the possibility of inferring socio-demographic information from written material in the Web 2.0.

1 INTRODUCTION

Social networks have emerged as a new source of information in modern sociology. Many areas are interested in the analysis of virtual communities as they are becoming more and more popular. There are many works based on analysing English social networks (Golder et al., 2007)(Mishne and Glance, 2006) (Gomez et al., 2008). As examples of applications of this analysis there are some related works: in *mass surveillance*, in *epidemiology* to help undestand how patterns of human contact aid the spread of diseases; to find new information or opinions for marketing strategies.

As far as we are concerned, the study of Spanish social networks is far behind the study of English social networks. We are using the data and the analysis described in section 2 to build a classification model which allows us to classify the user by origin, age and gender. Using the vocabulary of the users, we compare the performance of several classification techniques such as NaiveBayes, Decision Trees and Support Vector Machines. Additionally, we study the dependence on the user text length and we analyse different ways of selecting the most significant words for classification. Classification results are quite interesting when classifying by origin. Therefore, given the user text we are able to distinguish where are users from. Advances in this kind of classification may allow to support some Web mining tasks such as suplantation of identity or even user identification.

In this study, we have exclusively performed an statistical analysis of the forum contents. For a syntactic or semantic analysis, first it would have been necessary to convert the text from chatspeak to standard Spanish.

This paper is organised as follows. Next section reports details on the data collection and some statistics. Section 3 explains the preprocessing and selection of the data. Section 4 briefly describes the classification techniques used and shows the results together with some experiments regarding the classification performance depending on the user text length. Finally, the last section presents the conclusions of this study.

2 SPANISH DATA COLLECTION

The social network community selected for this study is the Spanish speaking community of MySpace (www.myspace.com) and, more specifically, the study focus on forums created by this community. MySpace was created in 2003, and has become the social network with the highest number of registered users, with roughly 148 million of actives users (Fumero and Garcia Hervas, 2008). In the spring of 2007, MySpace launched a Latin version for Hispanic users that live in USA and other for Latino American public in general. With this launching it was already possible to use MySpace with the Spanish language. In Spain was officially presented in June of 2007, although the months before it was already possible to use MySpace in Spanish in a beta version. Nowadays, according to recent studies (U. McCann, 2008) MySpace is also

R. Costa-jussà M., E. Banchs R. and Codina J. (2010).
WHERE ARE YOU FROM? - Tell me HOW you Write and I Will Tell you WHO you Are.
In *Proceedings of the 2nd International Conference on Agents and Artificial Intelligence - Artificial Intelligence*, pages 406-410
DOI: 10.5220/0002580504060410

the leader social network in Spain, with a percentage of 34% of users that connect in a social network. The dataset contains the entire amount of forum discussions in Spanish available at MySpace at May 16th, 2008. The total amount of retrieved data was 1.7GB. The oldest comments in the dataset where published on December 13, 2007 and the most newest on the very same day the data was collected.

2.1 Dataset Basic Statistics

The dataset contains about 300000 comments produced by about 23340 distinct users. The comments are divided into approximately 25,000 threads. The longest forum thread has more than 15,000 comments, which represents more than 5% of all comments observed in the study and the most prolific user has more than 8500 comments.

The number of posts per thread and the number of posts per user follow a log-normal distribution with a heavy tail that follows a power-low with cut-off (Kaltenbrunner et al., 2009)

Males and Females follow a similar population pyramid with slight differences (the most common age for males is 18 and 17 for females, the median is 22 years old for males, and 20 for females)

There are 18 countries with more than 50 users, Starting with Mexico (30% of all users), Spain (4000 users), the third are the 3500 users from the US while a 10% of the users does not specify the country they belong to.

3 PREPROCESSING AND SELECTION OF THE DATA

Given the original Spanish data, we found out that there were many words affected by encoding inconsistencies. This problem is specially problematic in the Spanish language given the many accentuated words and special characters as *ñ*. Therefore, we had to perform a manual preprocessing to normalize all characters that were affected.

The data forums contained many users from over 20 different countries. Given that there were not many users for many countries, which generated much sparseness in the data, we decided to make a selection of countries. In order to identify users by origin, we have discarded those countries which Spanish is not the official language, as the Spanish speakers in this area may be not native or immigrants from any country. Additionally, we selected countries with at least over 30 users who had written more than 100 words in the forums. The last requirement was arbitrarily chosen to ensure enough training data to train the classifier. This filter kept the following countries: Mexico, Spain, Argentina, Chile, Colombia, Peru and Venezuela.

Analyzing the age distribution, it shows that there is an abnormal number of users older than 80 years. This can be because either they did not fill in the birth date correctly or they just invented. We discard all users older than 80 years.

4 EXPERIMENTS WITH STANDARD CLASSIFICATION TECHNIQUES

Recent studies using the English language show that a number of stylistic and content-based indicators are significantly affected by both age and gender (Argamon et al., 2007).

In this study, we propose to analyse all the contributions to the different forums of one user. We study if the available information (bag of words, number of comments, discussions, starting threads and typed characters) can be useful to determine the user's country, gender and age.

Given the seven countries mentioned in the previous section, we made a wider classification into 3 categories: Central America, South America and Spain.

Given the variety of ages (from 14 to 80), we also made a wider classification into 3 categories: 14 to 18 years, 19 to 24 years and more than 25 years.

Finally, the gender has 2 categories: male or female.

In order to classify, we used the open source software WEKA [1]. Our attributes to classify were words and by default we selected the most relevant ones by means of the standard TF-IDF weighting (Salton and McGill, 1983). Evaluation was done by an standard *10*-fold cross-validation.

Figure 4 shows the user distribution in each category. This distribution allows us to set a baseline result in classification. Regarding the gender, the baseline reference is around 51%. Regarding the origin, the baseline reference is around 45%. Finally, regarding the age, the baseline reference is around 35%.

Given the proposed classifications and their categories, we want to analyse the following:

1. Which are the features that allow to improve the classification?

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

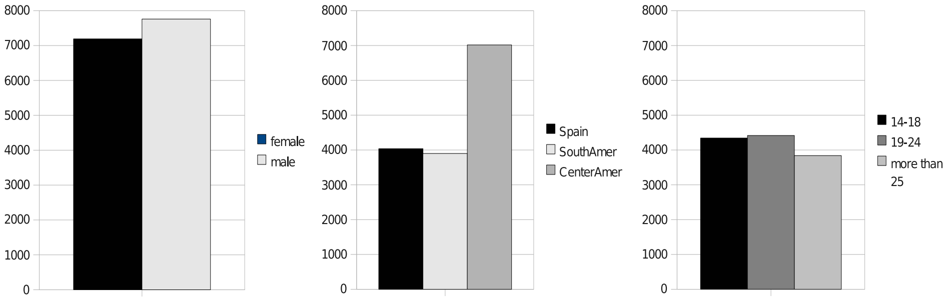

Figure 1: Distribution of the users among the different classifications: gender, origin and age.

2. Which aspects of the users (origin, age or gender) can be better derived from the way they write?
3. Which classifiers achieve the best performance among: Naive Bayes, Support Vector Machines and Decision Trees?

Next subsections are dedicated to answer and discuss the above questions.

4.1 Analysis of the Features

Together with the user texts, we have access to some additional information that we can use for the classification. We have available for each user the number of comments, discussions, starting threads and the number of typed characters. Additionally, in case of classifying the origin, we can use the gender and age as explicative attributes by themselves. Similarly, in case of classifying the age, we can use the origin and gender. Finally, in case of classifying the gender, we can use the origin and age.

Table 1 shows the results of using the combination of all the attributes named above, and the influence of discarding each one of them. *ZeroR* is the approach using bag of words. We can see that, in general, these attributes do not have any valuable information for classification. Note that discarding most of them does not change the quality in classification. However, there are some interesting conclusions: using the user age to find out the user origin helps to increase the classification results and using either the gender or the origin to find out the user age also helps to increase the classification results. In terms of gender classification, we were not able to find any contribution from any attribute.

4.2 Text Length

In order to train a classifier based on the user-generated texts, one might think that a minimum of user activity (in terms of written words) is needed. That is why, we study this effect in classification by discarding users having less than a given number of written characters. Beforehand, we experimented whether the distribution of categories inside each classification varied. No significant variation was observed.

Table 1: Percentage of correctly classified instances using different features.

Attributes	Origin	Age	Gender
ZeroR	45.24	35.02	51.24
ALL	47.24	45.68	51.24
# comments	47.24	45.68	51.24
# discussions	47.24	45.68	51.24
# threads	47.24	45.68	51.24
# chars	47.24	45.68	51.24
origin	-	**37.99**	51.24
age	**45.24**	-	51.24
gender	47.24	**43.80**	-

Figure 2 shows the performance in gender, origin and age classication given different text length constraints and using a simple classifier based on Naive Bayes. For the gender and the origin classification, results are quite irregular and it seems there is no clear correlation between the performance in classification and the text length used to train the classification. However, for the age classification, it seems that if the text is very short, the performance in classification falls.

Regarding the quality in classification, we can observe that simply using Naive Bayes we get quite good results when classifying by origin and age: +12% and +13%, respectively. The gender classification gives pretty bad results, which may be caused by different kind of reasons such as user falsification of the gender or simply because men and women tend to use similar words.

4.3 Classifier Experiments and Attribute Selection

At this point, we keep only the challenge of origin classification. However, experiments could be done

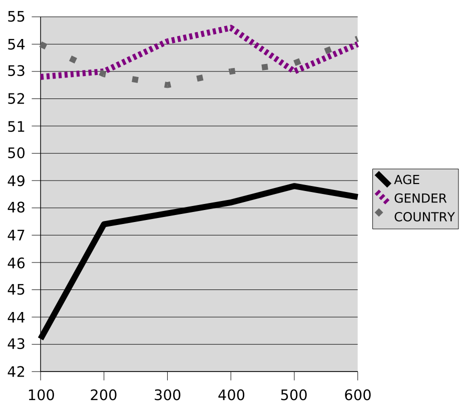

Figure 2: Classification performance depending on the text length. The most relevant 1000 words, according to TF-IDF, were considered.

similarly for the age classification which performed quite well in the above section.

Here, we experiment with different classifiers. As mentioned before, our attributes are words and we consider the top 1000, 4000 or 15. The two formers were selected using the standard TF-IDF. The latter was selected over the second one and using simply a selection of the ones that had more impact for classifying.

Table 2 show the difference in performance. We observe that the best results are obtained when using Support Vector Machines and we reach up to 71% of correctness in classification which is quite interesting. Regarding the attributes, it is better to use as many words as possible. Unfortunately, it was computationally very expensive to make the experiment with CART and 4000 words (it could not be run with 4G of RAM).

Table 2: Percentage of correctly classified instances in origin classification using (1) different classifiers and (2) different attributes or words.

Words	ZeroR	NaiveBayes	SVM	CART
1000	45.24	52.01	69.75	59.83
4000	45.24	54.64	**71.10**	-
15 (4000)	45.24	59.65	59.83	59.62

Finally, Table 3 shows the confusion matrix for the best result in Table 2 over test sets.

Table 3: Confusion matrix for the best result from Table 2. CA stands for Central America and SA stands for South America.

	Spain	CA	SA
Spain	1610	244	228
CA	356	2439	489
SA	296	485	1112

5 CONCLUSIONS

This work constitutes a preliminary study about the feasibility of predicting the origin, the age and the gender of a social network user, given his/her posts in one or several forums. The study was conducted with MySpace forums in Spanish language, aiming at classifying and extracting demographic information along with age and gender differences. Different supervised classification techniques were evaluated, and the very simple bag-of-words approach was used as feature-space model.

The preliminary results reported in this study suggest that, for the cases of region of origin and age group, it is possible to generate predictions up to some extent well beyond random guess estimations. This means that the way the users express themselves in the Web provides some valuable information about their basic socio-demographic characteristics, at least for the case of Spanish speaking communities. We consider these findings very valuable for some important endeavors, such as automatic detection of identity supplantation, which can be used for protecting online community members by the early detection of possible online criminal activity.

For future work in this research line we intend to improve classification performance by means of a more complex text analysis. In this way, we will work on implementing some text normalization preprocesses in order to improve the quality of the texts, as well as extracting and evaluating some linguistically richer attributes by using morpho-syntactic and semantic analysis.

ACKNOWLEDGEMENTS

The authors would like to thank Jens Grivolla and Andreas Kaltenbrunner for their advice. This work has been partially funded by Barcelona Media Innovation Center and the Spanish Ministry of Education and Science through the *Juan de la Cierva* research program.

REFERENCES

Argamon, S., Koppel, M., Pennebaker, J. W., and Schler, J. (2007). Mining the blogosphere: Age, gender and the varieties of self-expression. *First Monday*, 12(9).

Fumero, A. and Garcia Hervas, J. M. (2008). Social networks. contextualizing the phenomenon of web 2.0. *TELOS. Cuadernos de Comunicacion e Innovacion*, (76). (in Spanish, online).

Golder, S. A., Wilkinson, D., , and Huberman, B. A. (2007). Rhythms of social interaction: Messaging within a massive online network. In *In 3rd International Conference on Communities and Technologies (CT2007)*.

Gomez, V., Kaltenbrunner, A., and Lopez, V. (2008). Statistical analysis of the social network and discussion threads in slashdot. In *In WWW08: Proceeding of the 17th international conference on World Wide Web*, pages 645–654, New York.

Kaltenbrunner, A., Bondia, E., and Banchs, R. E. (2009). Analyzing and ranking the spanish speaking myspace community by their contributions in forums. In *WWW '09: Proceeding of the 18th international conference on World Wide Web*. accepted.

Mishne, G. and Glance, N. (2006). Leave a reply: An analysis of weblog comments. In *In WWW2006, 3rd Annual Workshop on the Weblogging Ecosystem*, Edinburgh.

Salton, G. and McGill, M. J. (1983). *Introduction to modern information retrieval*. McGraw-Hill.

U. McCann (2008). Internacional Social Media Rearch. Wave 3. published online.