Author Profiling Using Corpus Statistics, Lexicons and Stylistic Features

Notebook for PAN at CLEF-2013

Maria De-Arteaga, Sergio Jimenez, George Dueñas, Sergio Mancera and Julia Baquero

Universidad Nacional de Colombia, Bogotá, Colombia
[mdeg|sgjimenezv|geduenasl|samanceran|jmbaquerov]@unal.edu.co

Abstract. This paper describes our participation in the 9th PAN evaluation lab in the author profiling task. The proposed approach relies on the extraction of stylistic, lexicon and corpus-based features, which were combined with a logistic classifier. These three sets of features contain pairwise intersections and even some features that belong to all categories. A comprehensive comparison of the contribution of several feature subsets is presented. In particular, a set of features based on Bayesian inference provided the most important contribution. We developed our system in the Spanish training corpus, once developed it was used, with minor changes, for the English documents, too. The proposed system was ranked 6th in the official ranking for Spanish documents among 17 submitted systems. This result shows that our approach is meaningful and competitive for predicting demographics from text.

Keywords: author profiling, gender prediction, age prediction

1 Introduction

Due to the large amount of textual information on the internet, it is now possible to carry out different research problems about the texts, either in connection with their authors, the registers involved, and the varieties of texts, among others. In the framework of the international conference, CLEF 2013 [L], we focused our study on the task of predicting demographic information about the authors from texts written in Spanish or English, by people of different age-range and gender.

In order to identify the author profile from written texts, the use of stylistic and content features is a common practice [A,B,D,E]. However others researchers prefer to focus only on the stylistic features [C]. The function words and part-of-speech are the main style-base features proposed for distinguishing the gender and age of the authors [A,B,C,D,E]. Another stylistic features included in these inventories are: the typical blog features [B], the grammatical and orthographic errors [A], the morphological, syntactic and structural attributes, and other stylistic characteristics extracted using the Linguistic Inquiry and Word Count (LIWC) program [D]. The most common measure employed is the frequency of each feature, normalized or not by the length of the document or other

criteria. Cheng et al. (2011) also includes some measures such Yules K, Simpsons D, Sichels S, Honores R and entropy.
The content-based features and the mechanism used for its selection also vary from one author to another. The extraction of corpus words for its comparison between the classes of interest [A,B], and the use of pre-established list of words [D,E] are the principal mechanism employed for the selection of this type of features.
In our study, each document is represented in a vector space, where each feature adds one unit to the dimension, including stylistic and lexicon-based attributes, relevant to distinguish the gender and age-range of the authors. Furthermore, we explore a new subset of features that involve the use of some statistics measures (corpus statistics features). These three subsets of features, as shown in Fig. 1, are intersected, and therefore some of them are located in more than one class. We used a machine-learning approach to build classification models to produce the predictions. The details of the task, documents and evaluation are presented in [L].
In the remainder of the paper, we begin with a description of the features (Section 2) and of the system used in this campaign (Section 3). Section 4 focuses on the main results of our work, while the final sections present the discussion and the conclusion that can be drawn from this study.

2 Features from Texts

The set of features extracted from each text contains components of one or more of the following categories: 'S' (style), 'C' (corpus statistics) and 'L' (lexicon). 1 shows a Venn diagram depicting the number of features extracted for each category combination. In the following subsections these features are described and the labels in Fig. 1 are used to clarify their categories, i.e. 'SL' for Style and Lexicon categories. Besides, the features in the 'C' category are presented separately by their supervised or unsupervised nature.

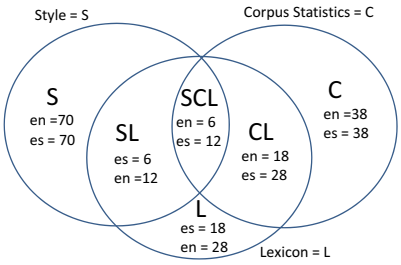

Fig. 1. Categories with their number of features by category and language

2.1 Unsupervised Corpus Statistics

This set of 6 features is built from statistics gathered from the training corpus, ignoring the demographic categories *age* and *gender* associated to each document. These corpus-based statistics use collection and document frequencies of the words in the entire training English and Spanish collections. The motivation for the use of document frequencies is to prevent very long documents from generating biased results.

IR features (2 'C' features). Using the $tf - idf$ term weighting approach used in the information retrieval field we obtained two features: $IDF(d) = \frac{\sum_{w \in d} idf(w)}{len(d)}$ and $TF.IDF(d) = \frac{\sum_{w \in d} tf(w,d) \cdot idf(w)}{len(d)}$, where $len(d)$ is the number of words in the document d, $idf(w) = log \frac{D}{df(w)}$, $df(w)$ is the number of documents where the word w occurs, D is the number of documents in the corpus and $tf(w,d)$ is the number of times that w occurs in the document d. $Tf - idf$ weight measures the informative character (for retrieval purpose) of the words given a particular document and the whole corpus. Thus, these features measure the density of that notion for each document.

Entropy (2 'C' features) measures the amount of information in a set of random variables, i.e. occurrences of words in a document. The probability of occurrence of a word is given by $P_f(w) \approx \frac{f(w)}{M}$ where $f(w)$ is the number of occurrences of w in the corpus, and M is the total number of words in the corpus. Alternatively, these probabilities can be obtained from document frequencies by $P_{df}(w) \approx \frac{df(w)}{D}$. Thus, the entropy of a document is given by $H_f(d) = \sum_{w \in d} P_f(w) \cdot \log_2 (P_f(w))$. $Hdf(d)$ is obtained with the same formula but using $P_{df}(w)$.

Kullback-Leibler (KL)-divergence (1 'C' feature) measures the information loss when a document probability distribution Q is used to approximate the "true" corpus distribution P. The probability Q for a word in a document is given by $Q_d(w) \approx \frac{df(w,d)}{len(d)}$. The corpus probability distribution P is given by $Pd(w) \approx \frac{f(w)}{\sum_{v \in d} f(v)}$. Thus, the KL-divergence of a document is given by $P_d \parallel Q_d(d) = \sum_{w \in d} P_d \cdot \ln \left(\frac{P_d(w)}{Q_d(w)} \right)$.

Cross entropy (1 'C' feature), similarly to the KL-divergence, compares P and Q measuring the ability of the former for predicting the latter. The cross entropy of a document is given by the following expression: $H(P_d, Q_d) = - \sum_{w \in d} P_d(w) \cdot \log_2 (Q_d(w))$.

2.2 Supervised Corpus Statistics

Unlike the previous set of features, this collection was built taking into account the *age* and *gender* of the authors of the training documents.

Gender score (2 'C' features). We developed the *gender score* (GS), a measure that aggregates the differences between the probabilities of a word w estimated in the corpus of documents written by males and females. Let $P_f(w|male) \approx \frac{f_{male}(w)}{M_{male}}$ be the probability of w estimated only in the corpus written by males, where $f_{male}(w)$ is the number of occurrences of w in the "male" subset of the corpus and M_{male} is the total number of words in that same subset. $P_f(w|female)$ is calculated analogously. Thus, GS is given by:
$GS_f(d) = \sum_{w \in d} (P_f(w|male) - P_f(w|female))$. GS_{df} is obtained using $P_{df}(w|male) \approx \frac{df_{male}(w)}{D_{male}}$ where $df_{male}(w)$ is the number of documents written by males where w occurred and D_{male} is the total number of documents written by males. Again, $P_{df}(w|female)$ is calculated analogously.

Bayes score (10 'C' features). We proposed a score for each one of the five demographic categories male, female, 10's, 20's and 30's using the Bayes theorem. These scores are given by the expression $BS_{f,cat.}(d) = \sum_{w \in d} P_f(cat.|w)$ having $cat. \in \{male, female, 10s, 20s, 30s\}$, $P_f(cat.|w) = \frac{P_f(w|cat.) \cdot P(cat.))}{P_f(w)}$ and $P(cat.) \approx \frac{Dcat.}{D}$. Similarly, $BS_{df,cat.}$ is obtained analogously but using probabilities subscripted by df. This way, we obtained 10 features from the 5 categories ($cat.$) and the 2 types of probabilities P_f and P_{df}.

Supervised KL-divergence (5 'C' features) can also be used to build supervised attributes. In this case, it measures the information loss when Q_d is used to predict the probability distribution of the subset of documents written by authors of the demographic category $cat.$This probability distribution is given by $P_{d.cat}(w) \approx \frac{f(w|cat.)}{\sum_{w \in d} f(v|cat.)}$, and the KL divergences are given by $P||Q_{cat.}(d) = \sum_{w \in d} P_{d.cat} \cdot \ln \frac{P_{d.cat}(w)}{Q_d(w)}$.

Supervised cross entropy (5 'C' features). As it can be expected, cross-entropy can also be calculated based on probability distributions of each individual demographic category. In this case, it measures how predictable Q is when using $P_{d.cat.}$. The equation to do so is $H(P,Q)_{cat} = -\sum_{w \in d} P_{d.cat}(w) \cdot log_2 (Q_d(w))$.

Supervised lexicon extraction using T-test (20 'CL' features). The Students *t-test*, frequently used in text mining, allows us to determine the most characteristic words of each demographic category by measuring the significance of the differences in the occurrences of the words on each category (gender) or between the category and the whole corpus (age). We used critical values in the T-table to build five lexicons, one for each gender and age range. This word

lists contain the words that have an absolute T-value greater than 2 for the given category, which are equivalent to around the three percent most relevant words of each demographic group. The construction method is different for gender and age categories. However, the following definitions are used in both cases: $S = \sqrt{P_f(w) - P_f(w)^2}$ and $S_{cat.} = \sqrt{P_f(w|cat) - P_f(w|cat)^2}$. In the gender T-function, as in the gender score, values greater than zero are characteristic of males and those less than zero are more often used by female. This value is given by $T_g = \frac{P_f(w|male) - P_f(w|female)}{\sqrt{\frac{S_{male}{}^2}{D_{male}} + \frac{S_{female}{}^2}{D_{female}}}}$.

Since the comparison cannot be made the same way when having three categories, a T-function was used for each age range, comparing the category with the general corpus. This function is given by the following equation, where *cat.* can only be an age range category: $T_{cat.} = \frac{P_f(w|male) - P_f(w)}{\sqrt{\frac{S_{cat.}{}^2}{D_{cat.}} + \frac{S^2}{D}}}$. This procedure provides 5 lexicons of words characterizing each demographic category.

2.3 Lexicon-based Features

The 5 lexicons built using T-test, as well as other pre-fabricated lexicons are used to generate 4 features for each one:

Lexical density (1 feature) is the ratio of content words to the total number of words. Ure, according to Johansson, introduced it in order to distinguish between words with lexical properties, and those without [F]. The concept of lexical density is developed by Halliday whose definition is "the proportion of lexical items to the total words" [G]. If $l_i(d)$ is the number of words that belong to the ith lexicon in document d, then $LD_i(d) = l_i(d)/len(d)$.

Weighted density (1 feature). The Spanish Emotion Lexicon [K] and the lists generated using T-test, provides a weight I_i to every word. Weighted density is given by: $WD_i(d) = \sum_{w \in d} I_i(w)/len(d)$. In lexicons that do not provide weights, 1 was used as weight.

Lexicon entropy (2 features). We calculate the entropy in relation to every lexicon using the following equation: $H_i(d) = \sum_{w \in d \cap l_i} P_f(w) \cdot log_2\left(P_f(w)\right)$. The fourth feature corresponds to the entropy calculated using $P_{df}(w)$.

The used lexicons and their sources are listed in Table 1. Manual preprocessing was applied to some lexicons by deduplicating and adding the gender variation for some Spanish words. Twenty CSL features result from the 5 T-test lexicons, the two entropy-related attributes of bad words, Internet and stopwords add 6 SCL features. Similarly, their densities add six SL features. For the remaining lexicons, their entropies generate CL features, and their densities L features. This generates on 22 L and 22 CL features for Spanish, and 8 L and 8 CL features for English.

Table 1. Websites where the lists of words were obtained (consulted in May 2013)

Lexicon Lang.	# Words	Source
Bad words en	458	urbanoalvarez.es/blog/2008/04/04/bad-words-list
Bad words es	2,147	rufadas.com/wp-content/uploads/2008/02/malsonantes.pdf
Cooking en/es	885/706	cocina.univision.com/recursos/glosario
Emotions en	3,487	eqi.org/fw.htm
Emotions es	2,036	www.cic.ipn.mx/ sidorov (6 lexicons)
Dictionary es	44,370	openthes-es.berlios.de
Dictionary1 es	14,720	dict-es.sourceforge.net
Internet es	1,567	www.techdictionary.com/chat_cont1.html
Internet en/es	689	pc.net/glossary (same lexicon used for both languages)
Legal en	1,011	www.susana-translations.de/legal.htm
Love-sex es	95	www.elalmanaque.com/El_Origen_de_las_Palabras
Sports es/en	709/642	www.wikilengua.org/index.php/Glosario_de_deportes
Stopwords en/es	127/313	NLTK Stopwords Corpus

2.4 Stylistic Features

The stylistic features are classified in three subsets: character-based, wordbased, and syntactic features. The character-based features contain 50 features, such as character density, uppercase or lowercase characters, letters, and special characters like the use of asterisk. All of them, except the letter count, have been used by other researchers for identifying the profile of the author. The word-based features include 11 measures for vocabulary richness, the length of words and density of hapax legomenas, dislegomenas, 3-legomenas until 5-legomenas. Syntactic features involve 9 attributes related to the regular punctuation such as colon, semicolon and question marks, among others. We also considered as stylistic features those obtained from lexicons such as *stopwords*, *Internet* and *bad words*.

3 System Description

The submitted system was built by extracting the features described in the previous section for each one of the first 20,000 documents in the English and Spanish training sets. That is, 166 features for English and 198 in Spanish; the difference is due to the different number of lexicons used on each language. For obtaining words from the character sequences in the documents xml tags were removed. Then each consecutive sequence of characters in the English or Spanish alphabet that was delimited by space, tab, enter or any punctuation mark, produced a word.

The statistics used in the calculation of the features that contain the label C were gathered using all the documents in the training set, i.e. 236,600 documents in English and 75,900 in Spanish. For each word w in the vocabulary we obtained: $f(w)$, $f_{male}(w)$, $f_{female}(w)$, $f_{10s}(w)$, $f_{20s}(w)$, $f_{30s}(w)$, $df(w)$, $df_{male}(w)$, $df_{female}(w)$, $df_{10s}(w)$, $df_{20s}(w)$ and $df_{30s}(w)$.

These datasets were used to train 4 logistic classifiers [J], one for each pair of target class (age and gender) and language. The used implementation was that

included in Weka v.3.6.9 [I]. The same feature extractor used in the training data was used to get features from the test documents. Then, the 4 classifiers provided the age and gender predictions for both languages.

4 Experimental Results

In this section the official results obtained by the proposed system for predicting authors age and gender in unseen documents are presented in Table 2. To assess the contribution of the different feature sets, additional experiments were carried out using a subset comprised of the first 20,000 documents from the training set. Each feature subset was evaluated using 10-fold cross validation and the average of ten different random folds is reported. Tables 3 through 6 show the results of these experiments.

Table 2. Official results obtained by our submitted system (accuracy)

Language	**Gender**	**Age**	**Total**	**Baseline**
Spanish (*es*)	0.5627	0.5429	0.3145	0.1650
English (*en*)	0.4998	0.4885	0.2450	0.1650

Table 3. Average accuracies for our 3 categories of feature sets an for all features

Feature Set	**Gender** *en*	**Age** *en*	**Gender** *es*	**Age** *es*
Statistic	0.8393(0.0005)	0.7860(0.0013)	0.8038(0.0007)	0.7866(0.0004)
Lexicon	0.5933(0.0010)	0.6198(0.0003)	0.6261(0.0007)	0.6446(0.0006)
Stylistic	0.5502(0.0012)	0.6048(0.0003)	0.5981(0.0008)	0.6336(0.0009)
All	0.8477(0.0023)	0.7809(0.0002)	0.8202(0.0013)	n/a

Table 4. Average accuracies for features obtained either using or not class attributes

Feature Set	Gender *en*	Age *en*	Gender *es*	Age *es*
Supervised	0.8432(0.0003)	0.7968(0.0006)	0.8155(0.0007)	0.7941 (0.0005)
Unsupervised	0.5487(0.0012)	0.6075(0.0006)	0.5990(0.0005)	n/a

Table 5. Average accuracy in subcategories in the "statistics" feature set

	Gender *en*	Age *en*	Gender *es*	Age *es*
Bayes	0.7951(0.0004)	0.7382(0.0015)	0.7696(0.0002)	0.7677(0.0003)
Cross_entropy	0.5527(0.0008)	0.5891(0.0006)	0.5376(0.0006)	0.5624(0.0004)
Kullback	0.5485(0.0005)	0.6034(0.0003)	0.5896(0.0005)	0.5952(0.0007)
tt Lexicons	0.5863(0.0006)	0.6204(0.0004)	0,6240(0.0005)	0.6377(0.0003)
Word_given_X	0.5416(0.0007)	0.6165(0.0003)	0.6152(0.0007)	0.5979(0.0003)

Table 6. Average accuracies for each lexicon (max. $\sigma = 0.0072$)

	Badwords	Cooking	Dictionary	Emotions	Internet	Legal	Love-Sex	Sports	Stopwords
Gender *en*	0.5288	0.5257	n/a	0.5267	0.5270	0.5305	n/a	0.5311	0.5304
Age *en*	0.5551	0.5673	n/a	0.5593	0.5697	0.5942	n/a	0.5945	0.5934
Gender *es*	0.5388	0.5041	0.5433	0.5282	0.5187	n/a	0.5361	0.5359	0.5335
Age *es*	0.5774	0.5625	0.5800	0.5709	0.5628	n/a	0.5707	0.5676	0.5701

5 Discussion

As shown in Tables 3-6, the best results for distinguishing gender were obtained in English and Spanish using all features, while the supervised attributes were better predictors for age-range. The age and gender were more appropriately identified using statistical features, although they were more suitable for typifying gender. The best statistic predictor in all cases was the features based on the Bayes theorem. The lexical and stylistic features were more useful to distinguish age than gender. Finally, the pre-established lists of words do not distinguish gender although they are useful for discriminating age.

6 Conclusions

We participated in the 9th PAN evaluation campaign with an author profiling system based on a set of features extracted from documents that were combined with machine learning. The features were designed in such a way that each one could contain at least one of the following components: stylometry, usage of pre-fabricated lexicon and corpus statistics. We developed this system for Spanish obtaining the 6th place in the official results among 17 participant systems. However, the same system adapted for English (replacing Spanish lexicons) performed poorly in unseen documents.
In a comprehensive comparison of different features we concluded that the features that provided the larger contribution were the ones obtained from corpus statistics. Particularly, the proposed score obtained using the Bayes theorem. To the extent of our readings such (or a similar) features have not been used in the past.

References

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

[B] Schler, J., Koppel, M., Argamon, S. and Pennebaker, J.: Effects of Age and Gender on Blogging. In: Proceedings. of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs (2006)

[C] Koppel, M., Argamon S. and Shimoni A.: Automatically categorizing written texts by author gender, Literary and Linguistic Computing 17(4), November 2002, pp. 401-412 (2008).

[D] Nguyen D., Smith N. and Ros, C.: Author age prediction from text using linear regression. In LaTeCH '11 Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pp 115-123, (2011)

[E] Cheng, N., Chandramouli, R. and Subbalakshmi, K.: Author gender identification from text. In: Digital Investigation, Vol 8, N 1, pp 78-88 (2011)

[F] Johansson, V.: Lexical diversity and lexical density in speech and writing: a developmental perspective. In: Lung Working Papers in Linguistics, Vol 53, pp 61-79 (2008)

[G] Halliday, M. A. K.: Spoken and written language. Geelong Victoria: Deakin University (1985)
[H] Thoiron, P.: Diversity Index and Entropy as measures of lexical richness. In: Computers and the Humanities, Vol 20, pp 197-202 (1986)
[I] Hall, M., Eibe F., Holmes, G., and Pfahringer, B.. The WEKA data mining software: An update. SIGKDD Explorations, 11(1), pp 10–18 (2009)
[J] le Cessie, S., van Houwelingen, J.C. Ridge Estimators in Logistic Regression. Applied Statistics. 41(1):191-201 (1992)
[K] Sidorov, G., Miranda-Jiménez, S., Viveros-Jiménez, F., Gelbukh, A., Castro-Sánchez, N., Velásquez, F., Díaz-Rangel, Suárez-Guerra, S. , Treviño, A., and Gordon J.. Empirical Study of Opinion Mining in Spanish Tweets. LNAI 7629-7630, pp 1-14 (2012)
[L] Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G. An Overview of the Traditional Authorship Attribution Subtask. CLEF (2013) (to appear)