ITALICA at PAN 2013: An Ensemble Learning Approach to Author Profiling

Notebook for PAN at CLEF 2013

Fermín L. Cruz[1], Rafa Haro R.[2], and F. Javier Ortega[1]

[1]Department of Languages and Computer Systems
University of Seville

[2]Zaizi
fcruz@us.es, rharo@zaizi.com, javierortega@us.es

Abstract This notebook discusses the approach to the Author Profiling task developed by the Italica group for PAN 2013. This system implements two different sets of classifiers which are combined later in order to build a final classifier that takes into account the decisions of the previous ones. The initial classifiers are focused on vector space representations of the documents as a bag of words and n-grams of POS tags and also on a set of stylistic features of the texts. The final classifier consists of a stacking schema that combines the other ones. This approach has obtained better results for the Spanish dataset than for the English dataset, probably due to the use of more detailed POS tagset in the former.

1 Introduction

In recent years, the WWW has suffered a social burst which, among other things, has resulted to an exponential growth of publicly available textual information generated by the users, mainly through blogs and social networks. The possibility of extracting valuable information from this huge amount of data has attracted the attention of researchers from different areas, focused on the possible ways of taking advantage of this wisdom of crows. In this context we can frame Author Profiling, which intends to extract as much information as possible about the author of a text by analysing the text itself. This task has many immediate applications in different domains, such as linguistics, forensics, user-centered design and commercial settings.

For the 9th evaluation lab on uncovering Plagiarism, Authorship, and social software misuse (PAN) competition, Author Profiling task deals with the classification of texts into classes of authors studying their sociolect aspect. In concrete, for the Author Profiling task in PAN, the problem consists in, given a document, determining the author's age and gender. Author Profiling has been addressed in some previous works. A corpus with more than 71K blogs classified by gender and nine age ranges is used in [8] with the goal of finding out if content independent features can categorize the authors. They take into account a set of features based on the textual content and others based on the writing style. Machine learning approaches for text classification [9] have been widely (and successfully) applied in domains where the classification depends on

content topics. In the other hand, the use of stylistic features as content independent features is also discussed in [11] and [1] in order to improve the classic vector space classifiers. From the point of view of linguistics, stylometry has been often applied to Authorship Attribution. A really good survey on the topic can be found in [5].

To develop their systems, participants were provided with a dataset consisting of 236K documents written in English and 76K documents in Spanish. Each document is composed of blog entries of a specific user. The gender (male, female) and the age (13-17, 23-27, 33-47) of the user are provided as well. The dataset is balanced in terms of gender class but highly unbalanced in terms of age class, where the amount of users from 13-17 age group is significantly smaller than the amount of users from the rest of age groups. Moreover, documents from authors who pretend to be minors have been included (e.g., documents composed of chat lines of sexual predators).

Our proposal is similar to [1], since it is based on supervised classifiers which takes into account both, content-based and writing style features. However, unlike the aforementioned proposals, we include a set of style-based metrics intended to measure the lexical richness of the authors as domain-independent features for the age and gender.

2 Our approach

Our approach is constituted by two components. First, we were interested in checking the performance of a classic text classifier based on a vector representation of the documents [9], whether it is at words level or using POS n-grams (with n=1, 2 or 3). On the other hand, we were interested in building a classifier based on stylistic features: use of accents and punctuation marks, lexical richness, capital letters, OOV words, ... Finally, both approaches are combined in a stacking schema.

2.1 Word and POS-based vector space classifiers

Following this idea, we have built classifiers based on vector space using a bag of words representation on one hand, and a bag of n-grams of POS tags on the other. For the bag of words representation input texts suffer 3 basic pre-processes: a conversion to lower cases, a tokenization based on Freeling [6] and the removal of stop words. We performed some experiments using just the word lemmas, but the results obtained made us opt for the use of the whole words, just as they appear in the texts.

We used Freeling in order to obtain the POS tags for the bag-of-POS representation. In the case of Spanish, the tags obtained with Freeling are from the PAROLE tagset [10]. These tags contain highly detailed information: for example, the tag DP3CS01 (assigned to word "mi") means that it is a possessive determiner in third person, of common gender, singular and with third person possessor. Although all this information could be useful in order to better characterize the language used by each type of authors, it would imply a lower statistical representation of the tags, more noticeable with trigrams. Therefore, we also built models based on a simplified version of the tagset such that, using the previous example, we would use just DP meaning that the word is a possessive determiner. For texts in English we use the Penn Treebank tagset whose tags are less detailed, hence we only built one representation of the documents.

We obtained different representations of the documents using n-grams with n=1, n=2 and n=3 of POS tags, hence each document is represented in six different ways for texts in Spanish (using the complete and simplified versions of POS tags) and three ways for texts in English. We used the same vector representation for the documents in all cases, choosing a maximum of 3K features represented by its tf-idf values. The selection of features was carried out using the chi-square correlation measure between the feature and the output classes. Therefore, we have different features chosen for the two classifications to be solved (for gender and age).

Once the vector representations are obtained, we trained the classifiers using WEKA [4]. We used the LibLINEAR [3] algorithm which implements a linear version of SVM. It is foreseeable to obtain better results using an SVM implementation with polynomial or radial kernel, but we cannot prove it because the training phase exceeded the available time.

2.2 Stylistic features

The usefulness of the lexical diversity of texts in the prediction of demographic features of the authors is studied in [7]. In contrast to the features based on tf-idf and n-grams, lexical diversity metrics are domain-independent and have a low computational cost. As other stylistic metrics, they are based on types/tokens relations which are intended to evaluate the diversity and amount of vocabulary contained in the texts. We use 17 normalized metrics of lexical diversity as stylistic features, as shown in Table1.

Metric	Formula	Metric	Formula
Nlemmas Density	$Nlemmas/Ntokens$	Root TTR	$Nlemmas/\sqrt{Ntokens}$
Lexical Density	$Nlex/Ntokens$	TTR Corrected	$Nlemmas/\sqrt{2Ntokens}$
Number of Words	$Ntokens$	Out-Of-Vocabulary Words	$NOOV$
Number of Sentences	$Nsentences$	Normalized OOV	$NOOV/Ntokens$
Avg. Number of Words	$Ntokens/Nsentences$	Capital letters	$Ncaps$
Lexical Diversity	$Tlex/Nlex$	Normalized Capitals	$Ncaps/Nletters$
Diversity G I	$Tgram/Nlex$	Vocals with accents	$Naccents$
Diversity G II	$Tgram/Ngram$	Normalized Accents	$Naccents/NVocals$
Diversity G Corrected	$Tgram/\sqrt{2Ngram}$	Punctuation marks	$Npuncs$
Token Ratio	$Nlemmas/Ntokens$	Normalized Punctuation	$Npuncs/Nchars$

Table 1. Metrics used in our proposal. Nlex is the number of lexical units (nouns, verbs, adjectives and adverbs). Ngram is the number of grammatical units (conjunctions, prepositions, pronouns and interjections). Tlex and Tgram are the number of distinct lemmas pertaining to lexical units and grammatical units, respectively. Nletters, NVocals and Nchars corresponds to the number of letters, vocals and characters, respectively.

2.3 Final ensemble

The combination of the classifiers was carried out as follows. First, the datasets of each language were partitioned into 10 sections. Using alternatively 9 different sections from each of the available vector representations, we trained 10 LibLINEAR classifiers which were applied to the remaining section in each case. The probabilities obtained by these classifiers for each class and for each vector representation were stored in a single output file for each dataset. This file is enriched lately with the stylistic features, obtaining a final file which will be used in the building of the final classifier. The final classifier is built with the rule-based algorithm JRip [2].

Our system performs the following steps in order to classify a document: first the document is processed with Freeling, whose output is used to build the bag-of-words and bag-of-POS models. Later we obtain the probabilities of the LibLINEAR-based classifiers applied to these models. These probabilities, in addition to the stylistic features, are used to build a final feature vector. Finally, we apply the JRip classifier.

3 Conclusions

According to the results achieved in the PAN competition, we have obtained the 3rd place for Spanish texts and 10th for English. For Spanish evaluation, our approach achieved 0.39 in overall accuracy, with 0.61 and 0.62 accuracy for gender and age, respectively. For English, we have obtained a total accuracy of 0.31 with 0.54 for gender and 0.59 for age. Since we do not know the dataset used for the evaluation in the competition, it is hard to make a comprehensive analysis of our proposal. Anyway, given the results achieved for Spanish texts, we can conclude that the use of a more complex set of tags for the POS-tagging in Spanish and also the use of the accents as a stylistic feature are two of the key points that could have made the difference.

References

1. Argamon, S., Koppel, M., Pennebaker, J.W., Schler, J.: Automatically profiling the author of an anonymous text. Commun. ACM 52(2), 119–123 (Feb 2009), http://doi.acm.org/10.1145/1461928.1461959
2. Cohen, W.W.: Fast effective rule induction. In: In Proceedings of the Twelfth International Conference on Machine Learning. pp. 115–123. Morgan Kaufmann (1995)
3. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: Liblinear: A library for large linear classification. J. Mach. Learn. Res. 9, 1871–1874 (Jun 2008), http://dl.acm.org/citation.cfm?id=1390681.1442794
4. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA Data Mining Software: An Update, vol. 11-1 (2009)
5. Holmes, D.: Authorship attribution. Computers and the Humanities 28(2), 87–106 (1994), http://dx.doi.org/10.1007/BF01830689
6. Padró, L., Stanilovsky, E.: Freeling 3.0: Towards wider multilinguality. In: Proceedings of the Language Resources and Evaluation Conference (LREC 2012). ELRA, Istanbul, Turkey (May 2012)
7. Roberto, J.A., Martí, M.A., Llorente, M.S.: Análisis de la riqueza léxica en el contexto de la clasificación de atributos demográficos latentes. Procesamiento del Lenguaje Natural 48, 97–104 (2012)
8. Schler, J., Koppel, M., Argamon, S., Pennebaker, J.W.: Effects of age and gender on blogging. In: AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs'06. pp. 199–205 (2006)
9. Sebastiani, F., Ricerche, C.N.D.: Machine learning in automated text categorization. ACM Computing Surveys 34, 1–47 (2002)
10. Tagset, W.P.: https://www.scss.tcd.ie/SLP/parole.htm
11. Wolters, M., Kirsten, M.: Exploring the use of linguistic features in domain and genre classification. In: Proceedings of the 9th Conference of the EACL. pp. 142–149. ACL, Stroudsburg, PA, USA (1999), http://dx.doi.org/10.3115/977035.977055