DAEDALUS at PAN 2014: Guessing Tweet Author's Gender and Age

Julio Villena-Román[1,2], José Carlos González-Cristóbal[3,1]

[1] DAEDALUS - Data, Decisions and Language, S.A.
[2] Universidad Carlos III de Madrid
[3] Universidad Politécnica de Madrid
jvillena@daedalus.es, josecarlos.gonzalez@upm.es

Abstract. This paper describes our participation at PAN 2014 author profiling task. Our idea was to define, develop and evaluate a simple machine learning classifier able to guess the gender and the age of a given user based on his/her texts, which could become part of the solution portfolio of the company. We were interested in finding not the best possible classifier that achieves the highest accuracy, but to find the optimum balance between performance and throughput using the most simple strategy and less dependent of external systems. Results show that our software using Naive Bayes Multinomial with a term vector model representation of the text is ranked quite well among the rest of participants in terms of accuracy.

Keywords: PAN, CLEF, author profiling, gender, age, user demographics, machine learning classifier, Naive Bayes Multinomial, term vector model.

1 Introduction

PAN[1] is a competitive evaluation lab on uncovering plagiarism, authorship and social software misuse, held as part of CLEF[2] conference. PAN 2014 offers three different main tasks: 1) plagiarism detection, 2) author identification and 3) author profiling. describes our participation at the PAN 2014 author profiling scenario [1]. We are a research group led by DAEDALUS[3], a leading provider of language-based solutions in Spain, and research groups of Universidad Politécnica and Universidad Carlos III of Madrid. We are long-time participants in CLEF, in many different tracks and tasks since 2003, and also in a previous edition of PAN [2].

The task is focused on author profiling, i.e., the problem to distinguish between classes of authors studying how language is shared by people, allowing to identify aspects such as gender, age, native language, or personality type. Specifically, the focus is on author profiling in social media messages. Author profiling is a problem of

[1] http://pan.webis.de/
[2] http://www.clef-initiative.eu/
[3] http://www.daedalus.es/

growing importance in different applications such as forensics, security, and marketing, for instance, to know the demographics of people that like or dislike their products, based on the analysis of blogs and online product reviews.

Given a document, the task is to determine its author's age and gender. Participants are provided with a training data set that consists of blog posts, Twitter tweets and social media texts written in both English and Spanish as well as hotel reviews written in English. Gender is a binary classification (male or female) and with regard to age, the following 5 classes are considered: 18-24, 25-34, 35-49, 50-64, >65. Differently to other CLEF labs, participants must not submit the results of their experiments using a provided test corpus, but else upload a software that runs within TIRA evaluation platform[4].

The idea behind our participation was to define, develop and evaluate a simple machine learning classifier able to guess the gender and the age of a given user based on his/her texts, which could become part of the solution portfolio of the company. We were interested to find not the best possible classifier that achieves the best accuracy, but to find the best balance between performance and throughput using the most simple strategy and less dependent of external systems. Our system and results achieved are presented and discussed in the following sections.

2 Our approach

The provided training data covers 1) four different types of corpus with presumably different language usage, 2) two different languages (English and French), and 3) two attributes to guess (gender and age). After several preliminary analysis using cross validation on the training corpora, we decided to build a machine learning classifier specifically trained for each combination of corpus-language-attribute, so 14 classifiers in all.

Table 1. Information of corpus

Corpus	Language	Authors	Texts
Blog	English	147	2 278
Review	English	4 160	5 452
Socialmedia	English	7 746	146 843
Twitter	English	306	201 432
Blog	Spanish	88	1 685
Socialmedia	Spanish	1 272	22 097
Twitter	Spanish	178	155 326

Table 1 shows the number of authors and texts for each training corpus. Given the heterogeneity of each corpus, where some have just a few long documents per author (such as in the review corpus) and others have many short texts per author (for

[4] http://www.tira.io/

instance Twitter corpus), we decided to design a two-level classifier: first, a document-oriented classifier, which guesses the gender and age of a given text, and then, an author-oriented classifier, which predicts the gender and age of a given user by aggregating the output of the first classifier for each text written by a given user. All corpora are equally balanced for gender and age, so the training is not affected by any class unbalance problem.

All 14 classifiers are trained with all texts for each combination of corpus, language and attribute. We used Weka 3.7 for performing our experiments and for developing our software to run in TIRA. Texts were tokenized using WordTokenizer to obtain a simple bag of words representation. The tokenizer allows to define split characters that are removed from the term vector space representation of the text. Besides the usual split symbols, spaces and some punctuation marks, we use some specific delimiters such as hashtags (#), usernames (@), emoticons, slashes, ampersands, question marks and hyphens that are used to separate words in SEO optimized URLs. Finally, as a high number of terms were low frequency numerals we decided to add numbers as well to help in normalization.

Regarding the document-oriented classifiers, a number of supervised algorithms were evaluated using cross validation, and finally, for its performance, we selected Multinomial Naive Bayes (NBM) classifier [3] with the default values for parameters. Different configuration parameters were tested to reach the conclusion that NBM was robust enough and other representations (bigrams, feature selection) were not adding additional value.

Results of this document-oriented classifier on training data using cross validation are shown in Table 2.

Table 2. Results for training data (document-oriented classification)

Corpus	Language	Gender	Age
Blog	English	0.8277	0.6485
Review	English	0.6852	0.3400
Socialmedia	English	0.6187	0.4445
Twitter	English	0.8726	0.7571
Blog	Spanish	0.8619	0.6660
Socialmedia	Spanish	0.6217	0.4439
Twitter	Spanish	0.8686	0.7598

The author-oriented classifier reads the output of the document-oriented classifier for each text written by a given author and predicts the gender and age using a simple voting strategy, i.e., returns the most frequent value among all texts, selected after some preliminary tests. Some other strategies were tested, such as a voting approach using a confusion matrix with different cost for each decision values, depending on the estimated accuracy for each class, but unfortunately we did not find any definite conclusion or improvement due to lack of time.

The final submission consists in a script written in PHP that reads the input test corpus and the output directory, and, using a loop, processes every file in the test corpus, reading all documents and creating two files in the arff format suitable for

Weka, one for gender and another one for age. Then Weka is called to obtain the predictions and then the output is aggregated to select the most frequent value that is chosen as the final output prediction for the author.

3 Results

The gender and age predictions have been evaluated as a classification problem, so accuracy measure over each class are reported. Results achieved by our software are shown in Table 3.

Table 3. Results for test data (author-oriented classification)

Corpus	Language	Gender	Age	Both
Blog	English	0.6410	0.3974	0.3077
Review	English	0.6845	0.3143	0.2199
Socialmedia	English	0.5421	0.3581	0.1905
Twitter	English	0.5130	0.4156	0.2078
Average		0.5952	0.3714	0.2315
Blog	Spanish	0.5179	0.4643	0.2321
Socialmedia	Spanish	0.5724	0.3622	0.1961
Twitter	Spanish	0.5444	0.5000	0.2667
Average		0.5449	0.4422	0.2317

In general, classifiers for Spanish achieve better results than classifiers for English, except for the case of blogs where English works better.

Although apparently gender attribute achieves a higher precision than age attribute, the classifier for gender is quite useless, as, taking into account that the range of values for the attribute is just two (male vs female), a random choice would achieve a 0.50 accuracy (assuming an equally balanced test corpus, the same as the training corpus). Thus classifiers for age outperform classifiers for gender in terms of lift (increment with regards to the random choice): for instance, 59% vs 50% for gender in English, 37% vs 20% for age in English (5 possible classes), etc.

Table 4 shows the comparison with other participants. This table shows, for each corpus, language and attribute, the maximum, minimum and average values, and the position of our software in the ranking of participants.

In general, we achieve average results just above the middle of the table, except for same cases were our software outperforms other participants, such as social media or reviews in English.

As it can be also noticed in the table, our results for Spanish are worse than the average for all participants in Spanish, though the approach is the same as for English. We do not have any explanation for this issue yet. However, we have a feeling that a stemming or lemmatization step should have been considered for Spanish, as inflection processes are important in this language and affect other tasks such as information retrieval or named entity recognition.

Table 4. Overall results.

Corpus	Language	Value*	Gender	Age	Both
Blog	English	Max	0.6795	0.4615	0.3077
		Min	0.5000	0.1795	0.0897
		Average	0.6117	0.3516	0.2326
		Ranking	3-4/7	2-3/7	1-2/7
Review	English	Max	0.7259	0.3502	0.2564
		Min	0.5012	0.0901	0.0451
		Average	0.6383	0.2879	0.1897
		Ranking	2/7	5/7	5/7
Socialmedia	English	Max	0.5421	0.3652	0.2062
		Min	0.5012	0.2355	0.1244
		Average	0.5285	0.3246	0.1750
		Ranking	1/7	3/7	4/7
Twitter	English	Max	0.7338	0.5065	0.3571
		Min	0.5065	0.1104	0.0584
		Average	0.5974	0.3766	0.2305
		Ranking	7/8	4/8	4/8
Blog	Spanish	Max	0.5893	0.4821	0.3214
		Min	0.4286	0.2500	0.1786
		Average	0.5112	0.4152	0.2366
		Ranking	3-4/8	3-4/8	4-5-6/8
Socialmedia	Spanish	Max	0.6837	0.4894	0.3357
		Min	0.5000	0.2191	0.1060
		Average	0.6144	0.3847	0.2325
		Ranking	7/8	5/8	6/8
Twitter	Spanish	Max	0.6556	0.6111	0.4333
		Min	0.5000	0.2222	0.1444
		Average	0.5736	0.4875	0.2889
		Ranking	5/8	5-6/8	6/8

* If there is more than one number in the ranking, it means a tie between participants

4 Conclusions and Future work

Results show that our quite simple approach using a two-level classifier composed of a document-oriented Naive Bayes Multinomial classifier with a term vector model representation of the text and then a voting strategy for predicting the author age achieves acceptable results in terms of accuracy. Despite of the difficulty of the task, results somewhat validate the fact that this technology may be already included into an automated workflow process for the first step towards social media mining and author profiling for supporting marketing activities.

However, in general, classifiers for gender (for all participants) are quite useless as they achieve a very low improvement over the random choice. Classifiers for age are worse in absolute accuracy but better in terms of lift with respect to the random choice. Obviously a different approach must be investigated to predict gender more robustly.

We already include a module for extraction user demographics in our portfolio of solutions [4], which tries to guess gender, age and user type (person or organization), using the information in the user public profile in Twitter, i.e., nick, full name and description, making no use of the texts written by that user. This module is based on distance among histograms using n-grams (character sequences) for each attribute to predict. Using internal evaluations, this software achieves good accuracy results for gender (over 70%) though lower for age.

Based on the results achieved in PAN, our initial idea to find a strategy that offers a good balance between performance and throughput using the most simple approach and less dependent of external systems gets validated and developing such classifier is within our immediate plans. In the short term, we plan to carry out some tests using our software for text classification [5], which is based on a hybrid algorithm [6] [7] that combines a statistical classification (currently based on kNN), which provides a base model that is relatively easy to train, with a rule-based filtering, which is used to post-process and improve the results provided by the previous classifier. We think that this combined strategy could provide improvements over these results based just on machine learning.

Regretfully, due to lack of time and resources, we have not been able yet to carry out an individual analysis by language, by corpus and a detailed analysis per class (confusion matrix) so we do not understand yet the effect of each component in the final result.

Specifically for the age attribute, we think that in a real business scenario, accuracy as defined in the task, i.e., a binary decision between right or not, could be somewhat relaxed using a cost matrix, considering that a miss classification between adjacent age ranges is less serious than between more distant ranges, specially for users who are near the end of the interval. So, we suggest to consider a modified evaluation metric that considers this cost matrix for future editions of PAN.

Acknowledgements

This work has been supported by several Spanish R&D projects: *Ciudad2020: Towards a New Model of a Sustainable Smart City* (INNPRONTA IPT-20111006), *MA2VICMR: Improving the Access, Analysis and Visibility of Multilingual and Multimedia Information in Web* (S2009/TIC-1542) and *MULTIMEDICA: Multilingual Information Extraction in Health Domain and Application to Scientific and Informative Documents* (TIN2010-20644-C03-01).

References

1. Francisco Rangel, Paolo Rosso, Moshe Koppel, Efstathios Stamatatos, and Giacomo Inches. Overview of the Author Profiling Task at PAN 2013. In Pamela Forner, Roberto Navigli, and Dan Tufis, editors, Working Notes Papers of the CLEF 2013 Evaluation Labs, September 2013. ISBN 978-88-904810-3-1.
2. Pablo Suárez, José Carlos González, Julio Villena-Román. 2010. A plagiarism detector for intrinsic plagiarism. Lab Report for PAN at CLEF 2010. CLEF 2010 Labs and Workshops Notebook Papers. 22-23 September 2010, Padua Italy. ISBN 978-88-904810-0-0.
3. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten. 2009. *The WEKA Data Mining Software: An Update*. SIGKDD Explorations, Volume 11, Issue 1.
4. Textalytics User Demographics v1.0 API. 2014. http://textalytics.com/core/userdemographics-info
5. Textalytics Text Classification v1.1 API. 2014. http://textalytics.com/core/class-info
6. Julio Villena-Román, Sonia Collada-Pérez, Sara Lana-Serrano, and José Carlos González-Cristóbal. 2011. *Método híbrido para categorización de texto basado en aprendizaje y reglas*. Procesamiento del Lenguaje Natural, Vol. 46, 2011, pp. 35-42.
7. Julio Villena-Román, Sonia Collada-Pérez, Sara Lana-Serrano, and José Carlos González-Cristóbal. 2011. *Hybrid Approach Combining Machine Learning and a Rule-Based Expert System for Text Categorization*. Proceedings of the 24th International Florida Artificial Intelligence Research Society Conference (FLAIRS-11), May 18-20, 2011, Palm Beach, Florida, USA. AAAI Press 2011.