Multilingual author profiling using SVMs and linguistic features

Roy Khristopher Bayot and Teresa Gonçalves
d11668@alunos.uevora.pt, tcg@uevora.pt

Department of Informatics, University of Évora, Portugal

Abstract. This paper describes various experiments done to investigate author profiling of tweets in 4 different languages – English, Dutch, Italian, and Spanish. Profiling consists of age and gender classification, as well as regression on 5 different personality dimensions – extroversion, stability, agreeableness, openness, and conscientiousness. Different sets of features were tested – bag-of-words, word ngrams, and POS ngrams. SVM was used as the classifier. *Tfidf* worked best for most English tasks while for most of the tasks from the other languages, the combination of the best features worked better.

1 Introduction

Author profiling has been of importance in the recent years. From a forensic standpoint for example, it could be used to determine potential suspects by getting linguistic profiles and identifying characteristics. From a business intelligence perspective, companies could target specific people through online advertising. By knowing the profile of the authors, companies would easily find what a specific group of people talk about online and device strategies to advertise to these people. They could also analyze product reviews and know what types of products are liked or disliked by certain people.

The growth of the internet where text is one of the main forms of communication is one of the reasons for a rising interest in author profiling. Through this growth, various corpora could be extracted, curated, assembled from different sources such as blogs, websites, customer reviews, and even twitter posts. Of course, this presents some problems. For example, people from different countries who use the same online platform such as Twitter or Blogger could behave differently in terms of text usage. This presents a difficulty in profiling. This work tries to take this difficulty into account by studying which kind of features are useful for different languages.

The aim of this work is to investigate the effect of syntactic information on author profiling in different languages. For this purpose, we used the dataset from PAN 2015 [12] since it has 4 different languages profiled almost in the same way on age, gender, and 5 personality traits - agreeability, conscientiousness, extrovertedness, openness, and stability. The four languages are English, Dutch, Italian, and Spanish.

There are three different sets of features that are investigated in this work. The first set is the bag of words features in the form of term frequency and term frequency inverse document frequency. The second set are word ngrams. And finally, we also study part of speech ngrams to see if information extracted from these features are useful for characterizing twitter users.

2 State of the Art

One of the first few works on author profiling is that of Argamon et al. in [1] where texts are categorized base on gender, age, native language, and personality. For personality, only neuroticism was checked. The corpus comes from different sources. The age and gender have the same corpus taken from blog postings. The native language corpus was taken from International Corpus of Learner English. Personality was taken from essays of pyschology students from University of Texas in Austin. Two types of features were obtained: content-based features and style-based features and Bayesian Multinomial Regression was used as a classifier. Argamon et al. had some interesting results where from the gender task, they were able to achieve 76.1% accuracy using style and content features. For age task with 3 classes, the accuracy was at 77.7% also using style and content features. For the native language task, the classifiers were able to achieve 82.3% using only content features. And finally, in checking for neuroticism, the highest obtained was 65.7% using only style features.

There has also been some research that uses datasets collected from social media. A particular example is that of Schler et al. in [14] where writing styles in blogs are related to age and gender. Stylistic and content features were extracted from 71,000 different blogs and a Multi-Class Real Winnow was used to learn the models to classify the blogs. Stylistic features included parts-of-speech tags, function words, hyperlinks, and non-dictionary words. Content features included word unigrams with high information gain. The accuracy achieved was around 80% for gender classification and 75% for age identification.

The work of Argamon et al. [1] became the basis for the work in PAN. It is an ongoing project from CLEF with author profiling as one of its tasks. The submission system has quite some interesting features. First, it now has three editions, one for each year. In every edition, a new aspect is being studied. Second, software is submitted to server with an evaluation system. And finally, the evaluation of the software involves a test data that is held out of the initial training data. The test data is only available to the organizers while the training data was distributed so that the software system could be made.

PAN currently has three editions. In the first edition of PAN [11] in 2013, the task was age and gender profiling for English and Spanish blogs. There were a variety of methods used. One set includes content-based features such as bag of words, named entities, dictionary words, slang words, contractions, sentiment words, and emotion words. Another would be stylistic features such as frequencies, punctuations, POS, HTML use, readability measures, and other various statistics. There are also features that are n-grams based, IR-based,

and collocations-based. Named entities, sentiment words, emotion words, and slang, contractions and words with character flooding were also considered. After extracting the features, the classifiers that were used were the following - decision trees, Support Vector Machines, logistic regression, Naïve Bayes, Maximum Entropy, Stochastic Gradient Descent, and random forests. The work of Lopez-Monroy in [4] was considered the winner for the task although they placed second for both English and Spanish with an accuracy of 38.13% and 41.58% respectively. They used second order representation based on relationships between documents and profiles. The work of Meina et al. [7] used collocations and placed first for English with a total accuracy of 38.94%. On the other hand, the work of Santosh et al. in [13] gave a total accuracy of 42.08% after using POS features for Spanish.

In PAN 2014 [10], the task was profiling authors with text from four different sources - social media, twitter, blogs, and hotel reviews. Most of the approaches used in this edition are similar to the previous year. In [3], the method used to represent terms in a space of profiles and then represent the documents in the space of profiles and subprofiles were built using expectation maximization clustering. This is the same method as in the previous year in [4]. In [5], ngrams were used with stopwords, punctuations, and emoticons retained, and then idf count was also used before placed into a classifier. Liblinear logistic regression returned with the best result. In [17], different features were used that were related to length (number of characters, words, sentences), information retrieval (cosine similarity, okapi BM25), and readability (Flesch-Kincaid readability, correctness, style). Another approach is to use term vector model representation as in [16]. For the work of Marquardt et al. in [6], they used a combination of content-based features (MRC, LIWC, sentiments) and stylistic features (readability, html tags, spelling and grammatical error, emoticons, total number of posts, number of capitalized letters number of capitalized words). Classifiers also varied for this edition. There was the use of logistic regression, multinomial Naïve Bayes, liblinear, random forests, Support Vector Machines, and decision tables. The method of Lopez-Monroy in [3] gave the best result with an average accuracy of 28.95% on all corpus-types and languages.

3 Experimental Setup

As mentioned in the introduction, three sets of features were placed under study: bag-of-words features, word ngrams, and part of speech (POS) tags ngrams. The same approach was used for all the tasks which was more or less straightforward: feature extraction, then use the features for either classification or regression.

After choosing the best features for each feature type, we did a fourth experiment combining all of the best features. The following subsections gives the details for the dataset, preprocessing, the features extracted, the learning algorithm, and the evaluation procedure.

3.1 Dataset

The data set was taken from PAN 2015 Author Profiling task [9]. It is composed of a set of tweets for 4 different languages – English, Dutch, Italian, and Spanish. It was decided to build a different model for each profiling element: age, gender, and the 5 different personality traits – extroverted, stable, agreeable, open, conscientious. There were 4 categories for the age classification - 18-24, 25-34, 35-49, and 50 and above but for Dutch and Italian, age classification is not possible because there was no data given. For personality traits, the values range from -0.5 to 0.5.

The number of given users varies for each language. There were 152 users for English, 34 for Dutch, 38 for Italian, and 100 for Spanish. Each user has a different number of tweets. The dataset is balanced based on gender.

3.2 Preprocessing

Processing the data was done through Python using the scikits-learn [8] library. For each language, xml files from each user are read. Then the tweets taken from each user are extracted and concatenated into one line to form one training example. The examples are then transformed by putting them all in lower case. No stop words were removed. Hashtags, numbers, mentions, shares, and retweets were not processed or transformed to anything else. The resulting file is used for feature extraction.

3.3 Feature extraction

Features were extracted from bag-of-words, word ngrams and POS ngrams.

For bag-of-words two different sets were built. The first is normalized term frequency (*tf*): terms are normalized by the number of terms in a training example; no terms were discarded. The second one is *tfidf* where terms were normalized by the inverse document frequency; to terms were also discarded.

Word ngrams were the second set of features examined. Counts of bigrams and trigrams were taken after preprocessing. No normalization was done and no terms were discarded.

Part of speech tags were also examined. They were extracted using a python wrapper to Schmid's TreeTagger program as detailed in [15]. Counts for unigrams, bigrams, and the combination of the two were used as features. The English parameter file for TreeTagger has 36 different tags, Italian has 38 tags, Dutch has 41 and Spanish has 75. No terms were discarded.

The number of features extracted for each set is given in Table 1.

3.4 Learning Algorithm

The learning algorithm used was Support Vector Machines [2] with a linear kernel and the parameter C chosen to be 1, which was the default setting. No parameter tuning was done in these experiments. The rationale for such is because the goal was to determine the effects of the features to the classification and regression.

Table 1. Number of features extracted

Feature	English	Dutch	Italian	Spanish
tf	26264	8569	12590	24688
tfidf	26264	8569	12590	24688
Word bigrams	104435	29648	36002	87313
Word trigrams	147577	39286	44051	122130
POS unigrams	35	39	38	68
POS bigrams	895	792	443	1999

3.5 Evaluation Procedure

Since the hold-out test data is not available (it can only be used in the PAN evaluation platform), we used a ten fold cross-validation procedure. After performing cross validation, a comparison between experiments of the same type was done using the Wilcoxon signed rank test [18] to check if the difference between features were significant or not. This test was used among other statistical tests because it allows the comparison between two experiments on the same data without making any assumptions on the distributions from which the set is drawn. A confidence interval of 95% was used.

Accuracy was the metric used for the age and gender task; for all problems involving personality mean squared error was used. The mean squared error is given in equation 1.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_{pred} - y_{actual})^2 \tag{1}$$

The objective is to get the maximum accuracy and the minimum squared error.

4 Experimental results

The following subsections detail, for each task, the results obtained for each set of features. The values in boldface indicate the results are statistically better when compared to the others.

4.1 Age Classification

The result for age classification is given in Table 2. *Tfidf* and word bigrams gave the best result among bag-of-words features and word ngrams respectively; for POS ngrams however, bigrams for English gave the best result while the combination of unigrams and bigrams gave the best result for Spanish. Looking at the statistical difference, only the results from bag-of-words features for English and word ngrams for Spanish were different from each other.

Table 2. Age classification results

Feature		English	Dutch	Italian	Spanish
BoW	tf	0.395			0.460
	tfidf	**0.691**			0.470
Text Ngrams	bigrams	0.638			**0.640**
	trigrams	0.553			0.460
POS Ngrams	unigrams	0.579			0.550
	bigrams	0.611			0.670
	uni+bi	0.593			0.700
combo		0.650			0.720

From the table we can see that *tfidf* features gave an accuracy of 69.1% for English; it is higher than the normalized term frequency and even the combination of the best features. For Spanish however, the combination of the best features gave the highest accuracy of 72.0%. Dutch and Italian don't have results since the corpus does not have any labels on this aspect.

4.2 Gender Classification

The results for gender classification are given in Table 3. *Tfidf* features gave better results than normalized term frequency. Word trigrams worked better than word bigrams for English and Dutch; on the other hand, word bigrams worked better for Italian and Spanish. Finally, the combination of POS unigrams and bigrams worked better for English, Italian, and Spanish, but POS unigrams worked better for Dutch. Looking at the statistical difference, only the results for word ngrams for Spanish, and bag-of-words for English and Spanish were different.

Table 3. Gender classification results

Feature		English	Dutch	Italian	Spanish
BoW	tf	0.511	0.450	0.617	0.550
	tfidf	**0.683**	0.517	0.750	**0.690**
Text Ngrams	bigrams	0.623	0.550	0.733	**0.700**
	trigrams	0.652	0.650	0.700	0.520
POS Ngrams	unigrams	0.659	0.642	0.733	0.770
	bigrams	0.659	0.567	0.733	0.750
	uni+bi	0.660	0.592	0.758	0.780
combo		0.671	0.658	0.758	0.800

From the table we can see that *tfidf* gave the highest accuracy result for English tweets with 68.3%; for the other languages, the combination of the features gave the best results: 65.8% for Dutch, 75.8% for Italian, and 80.0% for Spanish. Only Italian has a tie with another feature set, the combination of POS unigrams and bigrams.

4.3 Personality Regression

The average mean squared error across all five personality traits is given in Table 4. Generally speaking, most of the results are not statistically different from each other.

Table 4. Average of all the personality dimensions

Feature		English	Dutch	Italian	Spanish
BoW	tf	0.029	0.026	0.025	0.033
	tfidf	0.025	0.018	0.021	0.027
Text Ngrams	bigrams	0.027	0.020	0.023	0.029
	trigrams	0.028	0.023	0.024	0.032
POS Ngrams	unigrams	0.035	0.034	0.031	0.126
	bigrams	0.038	0.022	0.017	0.027
	uni+bi	0.040	0.022	0.018	0.029
combo		0.029	0.019	0.017	0.026

For the extroversion component, results within bag-of-words features are not statistically different from each other, on all languages; the same could be said for word ngrams on all languages. The only results that differ statistically is for Spanish using POS unigrams; it varies differently from that of both POS bigrams and the combination of POS unigrams and bigrams, with either of the two being the better result. Comparing the best results with the result with combined features, we also see that we don't gain anything.

I did not understand next sentence

For the stability component, results within bag-of-words features are not statistically different from each other on English, Dutch, and Italian; *tfidf* performed better than normalized term frequency in Spanish. Results within word ngrams are not statistically different from each other. For POS ngrams, results were not statistically different for English and Dutch; for Italian and Spanish however, POS bigrams and the combination of POS unigrams and bigrams gave better results and are statistically better than the results from POS unigrams alone. Finally, looking at the results for the combination of features, we do not really gain much.

For agreeability component, results within bag-of-words and word ngrams features are not statistically different from each other on all languages. For POS ngrams, only Spanish POS bigrams and the combination of POS unigrams and

bigrams gave better results and are statistically different from POS unigrams. Looking at the results for the combination of features, we can also see that we do not gain that much from the combination.

For the openness component, results with word ngrams features are not statistically different from each other on all languages. This is also the same for bag-of-words features except for Spanish where*tfidf* outperforms normalized term frequency. This is also the same for POS ngrams: all results are not statistically different from each other except for Spanish where POS bigrams and the combination of POS bigrams and unigrams gave better results than POS unigrams. Looking at the results for the combination of features, we can again see that the combination does not give better results.

Finally, for conscientiousness, results for bag-of-words features and word ngrams are not statistically different from each other. For POS ngrams we observe the same pattern where there is no statistical difference between each features for all languages except Spanish where POS bigrams and the combination of POS unigrams and bigrams gave better results than POS unigrams. Also, the combination of features does not improve the regression result.

5 Conclusions and Future Work

Comparing sets of features, *tfidf* generally works better than normalized term frequency, word bigrams also work better for most tasks than word trigrams, but for POS ngrams the best set of features is dependent on the task: .

enumerate which is better for each task

Looking at the written language, *tfidf* features worked best for all profiling tasks in English except conscientiousness regression. For Dutch, the combination of features gives better results for gender classification and extroversion regression; however, other features worked better for the other tasks: POS unigrams and bigrams for regression on stability, *tfidf* on agreeability and openness, POS unigrams and bigrams also for openness, and the combination of the best features for conscientiousness. For Italian , two types of features gave the best results: the combination of the best features and the combination of POS unigrams and bigrams. Finally, for Spanish the combination of all the best features works for most of the tasks: age and gender classification and extroversion and agreeability regression; POS bigrams worked best for stability and openness regression, while *tfidf* worked best for conscientious regression.

These results are by no means exhaustive; the conclusions need to be verified on other corpora and there are still many methods that could be explored. For instance, the preprocessing is quite minimal; more features could be extracted from the text such as number of links, retweets, hashtags, and mentions, length of tweets, ratios of uppercase to lower case characters, non-dictionary words, lexical diversity, emoticons, sentiment words, informative words and character ngrams. Furthermore, since Support Vector Machines with linear kernel was used, it would also be worth exploring other kernels as well as doing parameter tuning. Finally, a multi-dimensional classification approach can also be explored: instead of having two distinct classes for age and gender classification for instance, the

target classes could be the combination of the two classes; instead of having 5 different regressions on personality traits, it would be possible to take all five traits at once. This is worth exploring since the traits can be codependent.

explain how could you do that

You can go further on future work mentioning word2vec and cnn.

References

1. Shlomo Argamon, Moshe Koppel, James W Pennebaker, and Jonathan Schler. Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2):119–123, 2009.
2. Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.
3. A Pastor López-Monroy, Manuel Montes-y Gómez, Hugo Jair Escalante, and Luis Villaseñor-Pineda. Using intra-profile information for author profiling.
4. Adrian Pastor Lopez-Monroy, Manuel Montes-y Gomez, Hugo Jair Escalante, Luis Villasenor-Pineda, and Esaú Villatoro-Tello. Inaoe's participation at pan'13: Author profiling task. In *CLEF 2013 Evaluation Labs and Workshop*, 2013.
5. Suraj Maharjan, Prasha Shrestha, and Thamar Solorio. A simple approach to author profiling in mapreduce.
6. James Marquardt, Golnoosh Farnadi, Gayathri Vasudevan, Marie-Francine Moens, Sergio Davalos, Ankur Teredesai, and Martine De Cock. Age and gender identification in social media. *Proceedings of CLEF 2014 Evaluation Labs*, 2014.
7. Michał Meina, Karolina Brodzinska, Bartosz Celmer, Maja Czoków, Martyna Patera, Jakub Pezacki, and Mateusz Wilk. Ensemble-based classification for author profiling using various features. *Notebook Papers of CLEF*, 2013.
8. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
9. Francisco Rangel, P Rosso, M Potthast, B Stein, and W Daelemans. Overview of the 3rd author profiling task at pan 2015. In *CLEF*, 2015.
10. Francisco Rangel, Paolo Rosso, Irina Chugur, Martin Potthast, Martin Trenkmann, Benno Stein, Ben Verhoeven, and Walter Daelemans. Overview of the 2nd author profiling task at pan 2014. In *Proceedings of the Conference and Labs of the Evaluation Forum (Working Notes)*, 2014.
11. Francisco Rangel, Paolo Rosso, Moshe Moshe Koppel, Efstathios Stamatatos, and Giacomo Inches. Overview of the author profiling task at pan 2013. In *CLEF Conference on Multilingual and Multimodal Information Access Evaluation*, pages 352–365. CELCT, 2013.
12. Francisco Rangel, Paolo Rosso, Martin Potthast, Benno Stein, and Walter Daelemans. Overview of the 3rd Author Profiling Task at PAN 2015. In L Cappellato, N Ferro, J Gareth, and E San Juan, editors, *CLEF 2015 Labs and Workshops, Notebook Papers*. CEUR-WS.org, 2015.
13. K Santosh, Romil Bansal, Mihir Shekhar, and Vasudeva Varma. Author profiling: Predicting age and gender from blogs. *Notebook Papers of CLEF*, 2013.
14. Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, volume 6, pages 199–205, 2006.

15. Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the international conference on new methods in language processing*, volume 12, pages 44–49. Citeseer, 1994.
16. Julio Villena-Román and José Carlos González-Cristóbal. Daedalus at pan 2014: Guessing tweet author's gender and age.
17. Edson RD Weren, Viviane P Moreira, and José PM de Oliveira. Exploring information retrieval features for author profiling—notebook for pan at clef 2014. *Cappellato et al.[6]*.
18. Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, pages 80–83, 1945.