A Random Forest Approach for Authorship Profiling

Alonso Palomino-Garibay[1], Adolfo T. Camacho-González[1], Ricardo A. Fierro-Villaneda[2], Irazú Hernández-Farias[3], Davide Buscaldi[4], and Ivan V. Meza-Ruiz[2]

[1]Facultad de Ciencias
[2]Instituto de Investigaciones en Matematicas Aplicadas y en Sistemas (IIMAS)
Universidad Nacional Autonoma de Mexico (UNAM)
Ciudad de Mexico, Mexico
[3] Pattern Recognition and Human Language Technology,
Universitat Politécnica de Valencia
Valencia, Spain
[4]Laboratoire d'Informatique de Paris Nord, CNRS (UMR 7030)
Universite Paris 13, Sorbonne Paris Cité, Villetaneuse, France

Abstract In this paper we present our approach to extract profile information from anonymized tweets for the author profiling task at PAN 2015 [10]. Particularly we explore the versatility of random forest classifiers for the genre and age groups information and random forest regressions to score important aspects of the personality of a user. Furthermore we propose a set of features tailored for this task based on characteristics of the twitters. In particular, our approach relies on previous proposed features for sentiment analysis tasks.

Keywords: Author Profiling, Random forest, Random Forest Regression, NLP, Machine Learning.

1 Introduction

Authorship profiling exploits the sociolinguistic observations of particular spoken and written language that different groups of people use. However to extract important information about an author (*e.g. demographics, personality and cultural background*) just by analyzing raw text has a high potential number of applications from market research to forensics. From a marketing perspective recommendation systems which are vital part of today's Web can benefit of extract the profile dimensions of potential costumers to improve the way recommendations are performed. Moreover large corporations may be attracted to know what type of people like or dislike their products, based on analysis of blogs and online product reviews. From a forensic point of view authorship profiling can help to identify characteristics of crime perpetrators when there are many or few specific suspects to consider [1]

In this edition of the *PAN 2015 Author Profiling*, the task was formally defined as follows[1]:

[1] As described in the official website of the competition `http://pan.webis.de/(2015).`

> *This task is about predicting an author's demographics from her writing. Participants will be provided with Twitter tweets in English and Spanish to predict age, gender and personality traits. Moreover, they will be provided also with tweets in Italian and Dutch and asked to predict the gender and personality.*

Our approach proposes to use classifiers for the **age** and **gender** information and a set of regressors for the personality traits: **extroverted**, **stable**, **agreeable**, **conscientious** and **open**. In particular these traits are specified by a score. In this work we explore the use of Random Forest for both aspects of the task, classification and regression [3].

Our approach heavily depends on tailored features for the task. We have three types of features: lexical, twitter statistics and word list based . The lexical corresponds to features extracted over the whole vocabulary of the tweets. Statistic of the tweets count different aspects of the typical format of tweets; for instance the use of for mention of other users, or # for the marking of the topic of the tweet. The word list features correspond to total scores or frequencies of the use of terms within a tweet. For this type of feature we only consider specific terms from different word lists. An important part of these word lists is based on previous research on sentiment analysis. We explore the used of terms which determine degrees of polarity, irony or affect.

This paper is organized as follows: In the second section we give a complete description of the designed features for this task. In the third section we describe our methodology for authorship profiling. In the fourth section we describe the corpora provided by the PAN workshop 2015. In the fifth section we show the results, in particular we evaluate the performance of the system with accuracy metric.

2 Feature Engineering

Text representation is fundamental and indispensable for automatic information processsing, in our approach we extract a set of tailored features from a collection of tweets of a particular user. Although different speech communities might tend to write about different topics and in different ways, there are two types of features used for authorship profiling: content-based and style-based. The following list presents the used features:

1. **BOW/TF-IDF:**
   Based on the Vector Space Model, tweets are represented as a vector where each component is associated with a particular word from the corpus vocabulary. Typically, each component value is assigned using the information retrieval measure *tf-idf* this technique has been extensively used in text mining, information retrieval and NLP to classify text.

2. **POS (*Parts of speech*):**
   Unigram and bigrams of sequences of POS tags. These were obtained using the Core NLP Standford POS tagger (English and Spanish) [7], and the Tree Tagger (Italian and Dutch) [12].

3. **Irony detection words list [11]:**
   Irony is difficult to be defined, generally humor denotes this rhetorical device, structural ambiguity can be represented by the dispersion in the number of combinations

among the words that constitute humor examples [11]. For this feature, frequency and total score of words in tweets from an irony detection counter which uses a predefined word list where essential to match this event. Two dimension of the list use the counter factuality and the temporal compression.

4. **Sentiment polarity word list [8]:**
   For this feature we extracted the total score of positive and negative terms in tweets from predefined word list, all the occurrences were represented as a frequency vector.

5. **Sentiword word list [2]:** For this feature we use *SENTIWORDNET 3.0*, a well studied lexical resource to model the semantic orientation of sentiment classification and opinion mining applications, The total score of positive and negative terms in tweets from *SENTIWORDNET 3.0.* that are in users tweets are counted, for positive and negative instances. Translation for Spanish and Italian language support where crucial.

6. **Affect word list [14]:**
   The total score of affect terms in tweets from a word list. All the words from the user tweets that occurred in the list and have a greater or lower score of affect terms are counted into a matrix. This can purvey evidence of the personality of the user.

7. **Taboo word list:**
   Frequency of taboo words used in predefined list. Slang words are frequent in younger age groups, particularly this can be a remarkable feature that may show the type of personality of an author.

8. **Emoticons:**
   Frequency of emoticons used from predefined list. This feature can provide the type of personality as well as the age group of a user. All the occurrences of the terms of that match in the profiles are represented as a feature vector.

9. **Punctuation:**
   Frequency of punctuation signs from a predefined list. This can catch the type of discourse structure and semantics of a user.

10. **Links:**
    A frequency of domain links is helpful to match sites that contain interesting topics for the different demographic dimensions, if the tweet is repeated several times with a link this can be considered as a primary source of information.

11. **Tweets statistics:**
    This feature extract diverse types of statistics from tweets. Number of words, letters, capital letters, capital letter in initial position, numbers, lower cases, sentences. *RT* for retweets, for citations of *usernames*, and # for self defined *topic* of the tweet. Stylometric analysis is useful to identify gender and age groups [5].

Besides the previous engineered features we also tested with positive and negative frequency terms from [6] and a histogram of the Jaccard similarity coefficients among users tweets. Empirically we found that none of these features helped the for the task, since our metrics fall after being evaluated with this features.

Table 1 shows the final configuration of the features per language.

Feature	English	Spanish	Italian	Dutch
1	tfidf	tfidf	tfidf	tfidf
2	1gram	Bigram	Bigram	Bigram
3	Freq/Score		Freq/Score	
4	Score pos/neg	Score pos/neg	Score pos/neg	Score pos/neg
5	Score pos/neg	Score pos/neg	Score pos/neg	Score pos/neg
6	Socre	Score	Score	Score
7				Freq
8	Freq	Freq	Freq	Freq
9	Freq	Freq	Freq	Freq
10	Freq	Freq	Freq	Freq
11	Freq	Freq	Stat	Freq

Table 1. Features and configuration used per language

3 Approach

Our approach to authorship profiling relies in applying machine learning techniques to map text into categories. First we take the lexical corpora provided by *PAN-2015* and labeled according to a category in function of a profile or user. For instance, for author gender analysis we labeled as male or female each set of tweets. From the above proposed features we yield a document-term matrix, this means that each tweet was represented as a numerical vector in order to abstract features.

Then a supervised method computes classifiers and regressors based on the random forest algorithm, to the training examples. Finally the predictive ability of both (*classification and regression*) is tested on the testing data. We built two classifiers for English and Spanish (gender and age) and one for Italian and Dutch (gender). Additionally we created five regressors one per personality trait per language. Each classifier and regressor was independent from each other. Random forests have outstanding in recent years since the classification accuracy of this type of algorithms have outperformed SVMs and other machine learning algorithms in other knowledge areas for instance bio-informatics and computational biology creating classification methods for cancer diagnosis based on micro-array data [13]. We assume that this type of ensemble methods hold true for NLP tasks. The goal of ensemble methods is to combine the predictions of several base estimators built with a given learning algorithm in order to improve generalizability/robustness over a single estimator [9]. For this task we focused in averaging methods, which are learning algorithms that yield several estimators independently and then average their outcomes. Intuitively the averaging estimator is better than any single base estimators, as a result of reduced variance.

Briefly in Random Forests (both, regression and classification), each estimator in the ensemble is built from a bootstrap sample from the training set. When the algorithm splits a node during the generation of the decision tree, the chosen split is no longer the best split of all the features. Rather, the split that is selected is the best split of a random subset of the features. Due this randomness, the bias of the forest usually slightly increases but, due to averaging, its variance decreases, usually more than compensating for the increase in bias, finally this produces a better model [9].

The training was performed with Scikit-Learn, a library that provides a comprehensive suite of machine learning tools for Python. It extends this general-purpose programming language with machine learning operations: learning algorithms, preprocessing tools, model selection procedures and composition mechanisms to create complex machine learning work-flows [9].

3.1 Parameters

For both, regression and classification `n_estimators`which is the number of trees in the forest, if `n_estimators` is larger accuracy will increase, however this will increase the complexity to compute an prediction output. By the other hand if a lower amount of estimators is used the variance will reduce, but it will increase de bias of the model. Empirically we found that a good set up for classification of genre was: `n_estimators = 2000`.

4 Corpora

The corpora consists of tweets in four languages: English, Spanish, Italian, and Dutch every language has a collection of tweets from different users. The tweets were anonymized by removing the *username* information from the author and the mention to other *usernames*. The tweets as expected contain orthographic and typographic errors, colloquialisms, jargon and meta information such as *re-tweets* and link information. Not all the tweets were written by the author, for instance *re-tweets* and some tweets produced by automatic systems associated to the user. Both gender and age demographics were provided by the users answering an online test, however the personality trait scores were extracted using a personality test.[2]

The **gender** variable can take two values: *male* and *female*. The **age** variable four: *18-24*, *25-34*, *35-49* and *50-xx*. While the five personality traits are assessed by a score which goes from -0.5 to 0.5. Table 2 presents the sizes and number of tweets per user available in the training corpora provided by the organizers of the task [10].

[2] Based on website: `http://your-personality-test.com/`

Language	Number of users	Tweets per user
English	152	100
Spanish	100	100
Italian	38	100
Dutch	34	100

Table 2. Length of corpus per language

5 Results

Using a cross validation setting over the corpora we evaluate the performance of our system as follows. For **gender** and **age** we report *F1-score* and root mean square error (*RMSE*) for the personalities traits.

Trait	English	Spanish	Italian	Dutch
Gender	0.706	0.750	0.773	0.765
Age groups	0.612	0.465	N/A	N/A
Extroverted	0.023	0.024	0.018	0.014
Stable	0.041	0.036	0.025	0.027
Open	0.018	0.025	0.025	0.014
Conscientious	0.021	0.023	0.013	0.012
Agreeable	0.021	0.020	0.023	0.020

Table 3. Performance in training/development set. F-score for gender and age classification, and RMSE scores for personality traits.

pan15-author-profiling-test										
Language	GLOBAL	RMSE	Age	Agreeable	Both	Conscientious	Extroverted	Gender	Open	Stable
Dutch	0.6703	0.1595	NA	0.1598	NA	0.1787	0.1604	0.5000	0.1055	0.1928
English	0.5217	0.1749	0.4085	0.1572	0.2183	0.1526	0.1676	0.5000	0.1582	0.2392
Italian	0.6682	0.1636	NA	0.1463	NA	0.1553	0.1336	0.5000	0.1831	0.1997
Spanish	0.6215	0.1660	0.5114	0.1536	0.4091	0.1473	0.1729	0.8295	0.1530	0.2035

Table 4. Final results on test produced by the TIRA system [4].

6 Conclusions

In this paper we described our methodology for authorship profiling with PAN-2015 corpora. Author profiling has growing importance for national security, criminal investigations, and marketing research [1]. Our methodology uses random forests model for classification and regression. For this work we build a baseline system for the author profiling task that uses set of general features.

Our system presented some failures with the classification of the **gender** class which affected our performance. Additionally, we believe that the training of models was over-fitted by the number of estimators in both classification and regression Random Forest models.

For further research we plan to perform a better feature engineering by adding more specific features of content and style for the authorship and to implement a hyper-parameter optimization to tune the models.

Acknowledgments We acknowledge Rodrigo Sanabria contributions to the source code in the early stages of the project.

References

1. Argamon, S., Koppel, M., Pennebaker, J.W., Schler, J.: Automatically profiling the author of an anonymous text. Communications of the ACM 52(2), 119–123 (2009)
2. Baccianella, S., Esuli, A., Sebastiani, F.: Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining.
3. Breiman, L.: Random forests. Machine learning 45(1), 5–32 (2001)
4. Gollub, T., Stein, B., Burrows, S.: Ousting Ivory Tower Research: Towards a Web Framework for Providing Experiments as a Service. In: Hersh, B., Callan, J., Maarek, Y., Sanderson, M. (eds.) 35th International ACM Conference on Research and Development in Information Retrieval (SIGIR 12). pp. 1125–1126. ACM (Aug 2012)
5. Goswami, S., Sarkar, S., Rustagi, M.: Stylometric analysis of bloggers' age and gender. In: Third International AAAI Conference on Weblogs and Social Media (2009)
6. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 168–177. ACM (2004)
7. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D.: The stanford corenlp natural language processing toolkit. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 55–60 (2014)
8. Nielsen, F.Å.: A new anew: Evaluation of a word list for sentiment analysis in microblogs. arXiv preprint arXiv:1103.2903 (2011)
9. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)
10. Rangel, F., Rosso, P., Potthast, M., Stein, B., Daelemans, W.: In: Cappellato L., Ferro N., Gareth J. and San Juan E. (Eds). (Eds.) CLEF 2015 Labs and Workshops, Notebook Papers. CEUR-WS.org, (2015).
11. Reyes, A., Rosso, P., Veale, T.: A multidimensional approach for detecting irony in twitter. Language resources and evaluation 47(1), 239–268 (2013)
12. Schmid, H.: Improvements in part-of-speech tagging with an application t german. In: In Proceedings of the ACL SIGDAT-Workshop. Citeseer (1995)
13. Statnikov, A., Aliferis, C.F.: Are random forests better than support vector machines for microarray-based cancer classification? In: AMIA annual symposium proceedings. vol. 2007, p. 686. American Medical Informatics Association (2007)

14. Whissell, C., Fournier, M., Pelland, R., Weir, D., Makarec, K.: A dictionary of affect in language: Iv. reliability, validity, and applications. Perceptual and Motor Skills 62(3), 875–888 (1986)