



THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

## Author Profiling using Stylometric and Structural Feature Groupings

### Citation for published version:

Grivas, A, Krithara, A & Giannakopoulos, G 2015, Author Profiling using Stylometric and Structural Feature Groupings. in GJF Jones, L Cappellato, N Ferro & E San Juan (eds), *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015.*. CEUR Workshop Proceedings, vol. 1391, CEUR-WS, CLEF 2015: Conference and Labs of the Evaluation Forum Experimental IR meets Multilinguality, Multimodality and Interaction, Toulouse, 8/09/15. <<http://ceur-ws.org/Vol-1391/68-CR.pdf>>

### Link:

[Link to publication record in Edinburgh Research Explorer](#)

### Document Version:

Peer reviewed version

### Published In:

Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015.

### General rights

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

### Take down policy

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact [openaccess@ed.ac.uk](mailto:openaccess@ed.ac.uk) providing details, and we will remove access to the work immediately and investigate your claim.



# Author profiling using stylometric and structural feature groupings

## Notebook for PAN at CLEF 2015

Andreas Grivas, Anastasia Krithara, and George Giannakopoulos

Institute of Informatics and Telecommunications, NCSR Demokritos, Athens, Greece  
{agr, akrithara, ggianna}@iit.demokritos.gr

**Abstract** In this paper we present an approach for the task of author profiling. We propose a coherent grouping of features combined with appropriate preprocessing steps for each group. The groups we used were stylometric and structural, featuring among others, trigrams and counts of twitter specific characteristics. We address gender and age prediction as a classification task and personality prediction as a regression problem using Support Vector Machines and Support Vector Machine Regression respectively on documents created by joining each user's tweets.

## 1 Introduction

PAN, held as part of the CLEF conference is an evaluation lab on uncovering plagiarism, authorship, and social software misuse. In 2015, PAN featured 3 tasks, plagiarism detection, author identification and author profiling.

The 2015 Author Profiling task challenged participants to predict gender, age, and 5 personality traits (extroversion, stability, openness, agreeableness, conscientiousness) in 4 languages (English, Spanish, Italian and Dutch).

It featured quite a few novelties compared to the 2014 task. The addition of 5 personality traits to be estimated for the task, a change from 5 to 4 classes in the age estimation task, as well as a reduction in the size of the training dataset from 306 instances to 152 instances - user profiles.

In this paper we present an approach for tackling the author profiling task. In the next section the different steps of our approach are presented in details, while in section 3 the evaluation of the method is discussed.

## 2 Approach

For the author profiling task we proposed a coherent grouping of features combined with appropriate preprocessing steps for each group. The idea was to create an easily comprehensible, extensible and parameterizable framework for testing many different feature and preprocessing combinations.

We mainly focused on the gender and age subtasks as can be seen from the general approach taken towards personality traits, were we used the same features for all 5 different cases.

The architecture of the system we developed is portrayed in Figure 1. We will only sketch the outline of the system here, we will go into more details in the next sections.

The layers that can be seen correspond to the data structuring, preprocessing, feature extraction and classification steps that are carried out for the training and test cases. We follow a different preprocessing pipeline depending on the group of features we want to extract. We then combine the two groups, apply normalization and feature scaling and move on to the classification step where we train our model.

In the data structuring part of system we create a document for each user by joining all his tweets from the dataset.

This document is then preprocessed in the case of stylometric feature extraction. We initially remove all HTML tags found in the document and then we clear all twitter specific characteristics and tokens, such as hashtags, @replies as well as urls from the text. Using this cleaned form we then check for exact duplicate tweets and discard any if found.

We then extract structural features from the unprocessed document and stylometric features from the processed edition of the document. After concatenating these features together we normalize and scale their values, in order to avoid complications that can arise in the classification stage due to features with numeric values that differ a lot.

The last step, is the classification stage, where we train a Support Vector Machine or a Support Vector Machine Regression model depending on the subtask.

## 2.1 Features

In the tasks of Author Profiling and Author Identification many different types of features have been deemed important discriminative factors. In the same spirit as [5], we tried to group together features in a coherent way, such that we could perform suitable preprocessing steps for each group. Also, by grouping together features in such a way, it would be easier later on to split the task into separate classification subtasks and use a voting schema to obtain a final result.

In this work, we created two groups of features, namely the stylometric and structural features. The structural group of features aimed to trace characteristics of the text that were interdependent with the use of the twitter platform. Features such as counts of @mentions, hashtags and URLs in a user's tweets.

The stylometric group of features tried to capture characteristics of context that a user generates in a non automatic way. Different features were tested, such as tf-idf of ngrams, bag of ngrams, ngram graphs [1], bag of words, tf-idf of words, bag of smileys (emoticons), counts of words that were all in capital letters and counts of words of size 1 – 20.

Table 1 summarizes which of the features mentioned above were used for each subtask.

We based the stylometric aspect of our approach on trigrams since they capture stylometric features well and are more extensible to unknown text when a small training set has been used, comparing to a bag of words approach.

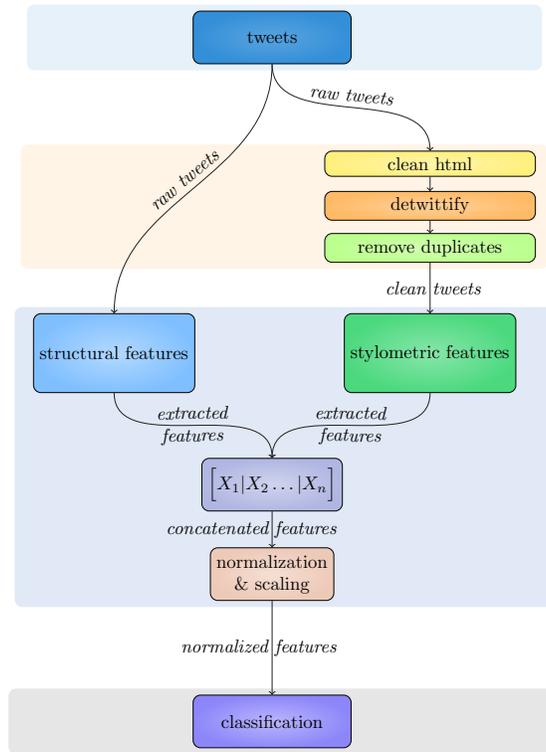


Figure 1: Architecture of the system

## 2.2 Preprocessing

Preprocessing is an important step which cannot be disregarded in this task. As texts are tweets, they contain specific information entangled in the text (hashtags, @replies and URL links). Therefore, an important decision involves deciding how to correctly deal with this bias.

Tweets also contain a large amount of quotations and repeated robot text, which may be structurally important but should be stylometrically insignificant.

In our approach, a different preprocessing pipeline was applied to each group of features as described above. There was no preprocessing done for structural features. Stylometric feature preprocessing encompassed removing any HTML found in the tweets, removing twitter bias such as HTML tags, @mentions, hashtags and URLs and removing exact duplicate tweets after removing twitter specific text. To elaborate a bit on removing twitter bias, @username and URLs were deleted, while hashtags were stripped of the hashtag character #.

In some approaches [2] that use tweets as a text source for classification, tweets are joined in order to create larger documents of text. For this task we joined all tweets for each user, however, it should make sense to try joining less texts and create more

Table 1: Features used for each subtask

Subtask	Group	Feature
gender	Stylometry	tfidf trigrams
	Structural	-
age	Stylometry	tfidf trigrams count word length
	Structural	count @ count hashtags count URLs
personality traits	Stylometry	tfidf trigrams count word length count caps word
	Structural	count @ count hashtags count URLs

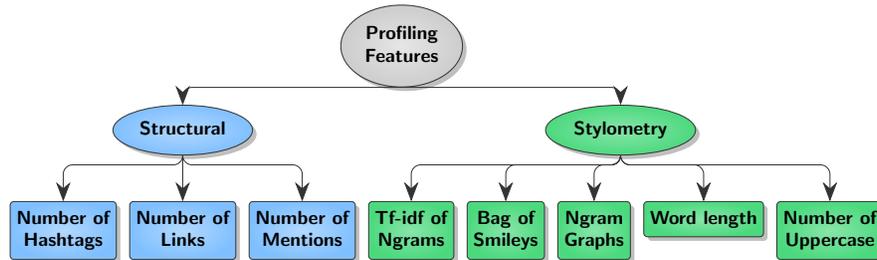


Figure 2: Groups of features

samples for each user, and then classify the user according to the label that has the majority of the predictions.

### 2.3 Classifiers

Regarding classification and regression, we used a Support Vector Machine (SVM) with a RBF kernel and a SVM with a linear kernel for the age and gender subtask respectively. In the case of the age subtask, we also employed the use of class weights inversely proportional to class frequencies since the distribution of instances in the classes was skewed. We used the implementations of the scikit-learn library [3] of the aforementioned machine learning algorithms.

Regarding the personality traits subtask, Support Vector Machine Regression (SVR) with a linear kernel was used.

For each subtask the features were concatenated and were then scaled and normalized. Scaling was performed in the features such that the values were in the range  $[-1, 1]$  with 0 mean and unit variance. Normalization was performed along instances so that each row had unit norm.

The above classifiers and combination of features were used for all languages of the challenge, namely English, Spanish, Dutch and Italian.

### 3 Evaluation

#### 3.1 Dataset

The Pan 2015 dataset featured less instances for training (152 users) than the earlier tasks in author profiling. The distribution of age and gender over the instances of the training set can be seen in Figure 3.

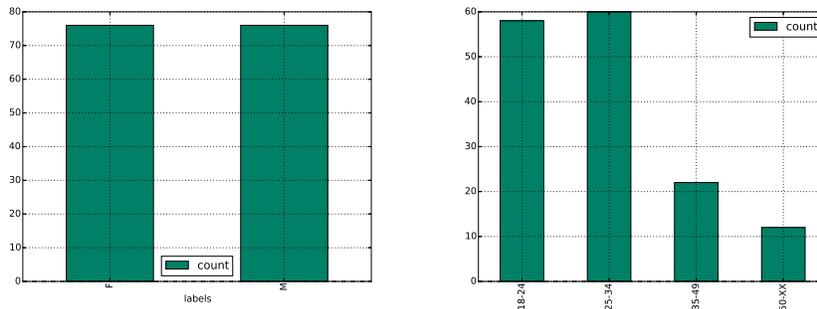


Figure 3: Age and gender distribution over training samples

#### 3.2 Performance Measures

In the context of Pan 2015 systems were evaluated using accuracy for the gender and age subtasks and average Root Mean Squared Error (RMSE) for the personality subtask. In order to obtain a global ranking the following formula was used.

$$\frac{(1 - \text{RMSE}) + \text{joint accuracy}}{2} \quad (1)$$

#### 3.3 Results

Our approach was in the top two approaches based on accuracy, regarding the gender classification subtask in all languages as can be seen in Figure 4. This fact hints that trigrams can capture gender information regardless of language and generalize well for datasets of this size.

However, results in Figure 5 show that our system performed less optimally in the case of age classification where more features that were considered helpful were used.

Using the scoring procedure described in Equation 1, our system scored 3rd overall in the over profiling task. An overview of the approaches and results for the author profiling task can be found in [4].

### 3.4 Future Work

In the context of our approach we will further evaluate the features used for the age classification subtask, in order to examine which of them are more useful and which actually deteriorate the performance of the approach on the test set. We will also develop a more sophisticated approach for personality trait identification, considering more specific features and preprocessing for each personality trait separately. Finally we will attempt to create more documents for each user by joining less tweets for each document and then arrive at a conclusion by using the average decision for all of the user documents. It will be interesting to see the impact of this approach on the results for each user.

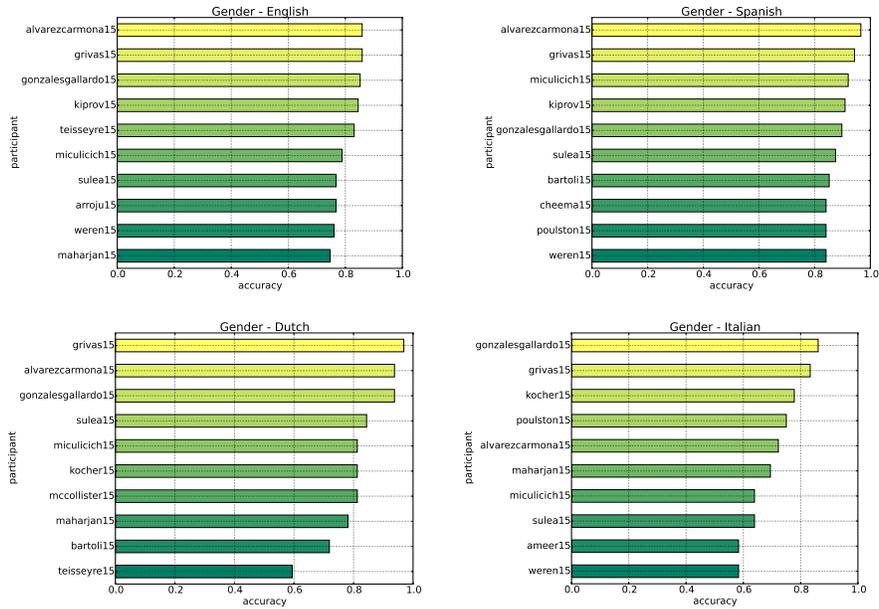


Figure 4: Top ten participants regarding gender prediction accuracy in all languages

## 4 Acknowledgments

This work was supported by REVEAL (<http://revealproject.eu/>) project, which has received funding by the European Unions 7th Framework Program for research, technology development and demonstration under the Grant Agreements No. FP7-610928.

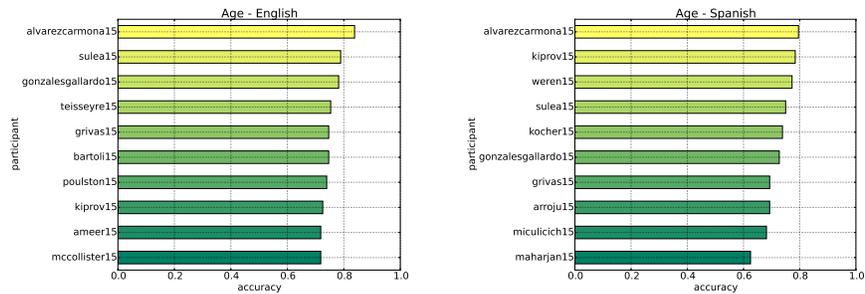


Figure 5: Top ten participants regarding age prediction accuracy in all languages

## References

1. Giannakopoulos, G., Karkaletsis, V., Vouros, G., Stamatopoulos, P.: Summarization system evaluation revisited: N-gram graphs. *ACM Trans. Speech Lang. Process.* 5(3), 5:1–5:39 (Oct 2008), <http://doi.acm.org/10.1145/1410358.1410359>
2. Mikros, G., Perifanos, K.: Authorship attribution in greek tweets using author’s multilevel n-gram profiles (2013), <https://www.aaii.org/ocs/index.php/SSS/SSS13/paper/view/5714>
3. 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)
4. Rangel, F., Rosso, P., Potthast, M., Stein, B., Daelemans, W.: Overview of the 3rd author profiling task at pan 2015. In: Cappellato L., Ferro N., Gareth J. and San Juan E. (Eds). (Eds.) CLEF 2015 Labs and Workshops, Notebook Papers. CEUR-WS.org (2015)
5. Stamatatos, E.: A survey of modern authorship attribution methods. *J. Am. Soc. Inf. Sci. Technol.* 60(3), 538–556 (Mar 2009), <http://dx.doi.org/10.1002/asi.v60:3>