Author Profiling using Complementary Second Order Attributes and Stylometric Features Notebook for PAN at CLEF 2016

Konstantinos Bougiatiotis and Anastasia Krithara

Institute of Informatics and Telecommunications National Center for Scientific Research (NCSR) "Demokritos", Athens, Greece {bogas.ko, akrithara}@iit.demokritos.gr

Abstract In this paper we present an approach for the task of author profiling. We propose a modular framework, extracting two main group of features, combined with appropriate preprocessing, implementing Support Vector Machines for classification. The two main groups we used were stylometric and discriminative, featuring trigrams on one hand and complementary-weighted Second Order Attributes on the other. We address the problem as a profile based problem creating target profiles and also grouping each user's tweets in the same document.

1 Introduction

PAN, held as part of the CLEF conference is an evaluation lab on uncovering plagiarism, authorship, and social software misuse. In 2016, PAN featured 3 tasks, author obfuscation, author identification and author profiling.

The 2016 Author Profiling task challenged participants to predict gender and age in 3 languages (English, Spanish and Dutch). It featured quite a few novelties compared to previous years. The first one is a big spurt in the size of the training dataset. Comparing to the 2014 task, where there were 152 user-profiles, this year the dataset comprises of 436 users, that is three times bigger. Moreover, we have an added class in the age estimation task, that of '65-XX', increasing the difficulty of the task. Finally, this year's task is focused on cross-genre identification. This means that the training documents were created from one genre (specifically Twitter) and the evaluation will be on other genres such blogs, social media etc.

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. In the final section, we draw conclusions and point out possible future work.

2 Proposed Method

In order to tackle the author profiling task we propose a semantically meaningful grouping of the features. We wanted to create a modular, extensible and parameterizable framework for testing different preprocessing, feature and classifier combinations. An overview of the methodology can be seen in Figure 1. The main steps involved in the architecture of the system are outlined as follows:

- 1. Preprocessing: Parsing, preprocessing and vectorization of the tweet, leading in a bag of words representation for each document-tweet.
- 2. Feature Extraction: Extracting discriminative and stylometry features from the processed text.
- 3. Classification: Train a Support Vector Machine classifier to distinguish between different age groups and genders using the extracted features from the collection

These steps are explained in details in the following subsections.



Figure 1. General Workflow of the system

2.1 Preprocessing

We start by applying a set of preprocessing procedures on the texts in our collection. In text mining, it is often assumed that words appear independently in a document and their order of occurrence is immaterial for the purposes of any information retrieval task at hand. Ultimately, this assumption leads to the bag of words representation of the document, according to which each document is represented as a multi-dimensional vector. This vector is populated with the counts of the different word appearances in each document, where each cell corresponds to a different word. The vector space is created by assigning a new dimension to each unique word in the document collection, that forms the vocabulary of the collection.

Still, before acquiring this representation of the tweets, the text must be cleared of Twitter specific information such as @replies, hashtags and URLs. In order to do so, all tweets are pipelined through a series of filters which remove any HTML found, URLs, numbers and unwanted punctuation marks. Moreover, the hashtags are stripped of their hashtag character #, leaving only the corresponding tag.

Afterwards, the texts are tokenized, splitted into words, and case folded, all letters reduced to lowercase. Thus, the resulting form of each tweet is a list of lowercase words, contained in the original tweet, after preprocessing. We also experimented with lemma-tization and stemming of the words [3], in order to decrease the term space by unifying variations of the same word, but to no avail with regards to improvement in results.

2.2 Feature Extraction

We focus on two different set of features, stylometric and discriminative. The stylometric set of features models the individual linguistic style of the users, that is existent in their corresponding written texts. These are ways of expressing oneself that can be mapped to user specific attributes such as age or gender. Features tested, as suggested by our previous work [1], were tf-idf of n-grams, bag of n-grams, bag of words and tf-idf of words.

The selected stylometric features are the frequencies of trigrams found in the tweets. Specifically for the age prediction task, we are interested only in the top 3000 most frequent trigrams found in the whole collection, while for the gender prediction the stylometric features didn't improve the results so they are not used.

Let it be noted here that structural features like counts of hashtags, URLs and other text characteristics, were also tested, but didn't enhance the outcomes and were abandoned.

The idea for the discriminative set of features is based upon the Second Order Attributes (SOA) [4]. The main concept is, at first, to associate the different words found in our collection with the different profiles, driven by the sense that many words are distinct per target profile. For the gender task, an example is that men talk more about "games" and women more frequently about "shopping", while for the age task, "pension" and "marriage" should be terms related more to middle-age groups rather than younger ones.

This is clearly depicted in Figure 2, where we can see terms like "dreamjob" that clearly denote a younger user, while others like "lol" that may denote a young agegroup but are not fully indicative of the age group. There also exist many discriminative terms regarding the gender subtask as shown in Figure 3.

After, having mapped the relations between each term and the target profiles for the task, we move on to create the document projections in the profile space. That is, we aim to create a vector for each document, where each cell would encompass the



Figure 2. Age descriptive terms



Figure 3. Gender descriptive terms

relationship of the document with the different target profiles, age or gender based. Documents are represented as the weighted aggregation of their terms' relations. This is depicted through the matrices multiplication in Figure 4.

Specifically, our approach introduces a few novelties in the way the final documents' profile relations are generated. First of all, we need to compute a value for each term in the vocabulary $V = [t_1, t_2, ..., t_V]$, indicating how relevant it is with each target profile $P = [p_1, p_2, ..., p_P]$. Let also $D = [d_1, d_2, ..., d_N]$ denote the collection of the documents-tweets. For computing the relation $t_{i,j}$, of term *i* with the profile *j*, one would normally take into account the documents containing the term *i* and at the same time belong in profile *j*. However, fueled by other studies [2],[7] and in order to counter any bias in our system due to skewed data (i.e. more training examples for one profile than another) we create each term-profile relation based on the complementary profiles in the collection. That is, instead of using training data from the profile *j*, we estimate the profile parameters using data from all profiles except *j*. This generates more robust estimates, because each one uses a more even amount of training data per profile, which lessens the bias.

Nonetheless, we would like to exploit the knowledge about the prior distribution of documents into profiles, that is $P(p_j)$. To do so, we incorporate a weighting term in the estimation procedure that is inversely proportional to the probability of the profile. The core idea, is that the rarer a profile, the more influence should the terms related to this



Figure 4. Documents' representations resulting from the aggregation of their terms

profile exhibit, in order to make up for the sparsity of the documents related with this profile. Finally, the term-profile estimate is:

$$t_{i,j} = \sum_{k:d_k \notin P_j} \log(1 + \frac{tf_{i,k}}{len(d_k)} * w_k)$$

where $tf_{i,k}$ is the term frequency in document k, $len(d_k)$ the document's length and w_k the weighting parameter of the class the document belongs to. We also apply a two-way normalization on the term parameters:

- Firstly, for each specific term over all the profiles, in order to emphasize the difference between the profiles related to the term and those who are not:

$$t_{i,j} = \frac{t_{i,j}}{\sum_{j=1}^{P} t_{i,j}}$$

- Secondly, for each profile over all the terms in the collection, so as to model the parameters per profile according to the value of the other terms also:

$$t_{i,j} = \frac{t_{i,j}}{\sum_{i=1}^{V} t_{i,j}}$$

Before moving on, let us clarify that using the complementary profiles as training data for each term-profile estimate, we get the opposite result of the previous depicted

examples. That is, the profile most related to a term should be the one with the minimum value in the vector $\mathbf{t} = [t_{i,1}, t_{i,2}, ..., t_{i,P}]$. That also stands for the document-profile vector $\mathbf{d}_{\mathbf{k}} = [d_{i,1}, d_{i,2}, ..., d_{i,P}]$, that will be created afterwards.

After projecting terms in the space of profiles, we build relationships between documents and profiles; these are the second-order attributes. Those are generated as shown in Figure 4, taking into consideration for each document only the terms present in it. This results in the document-profile vector d_k where each cell is a real value representing the relationship of the document d_k with a profile p_j . Lastly, while summing the term vectors, they are firstly weighted by the relative frequency of the term t_i in the current document. That is:

$$d_{k,j} = \sum_{i:t_i \in d_k} \frac{tf_{i,k}}{len(d_k)} \times \boldsymbol{t_i}$$

After the feature extraction procedure, we concatenate the features in a new feature vector containing all the generated features, as shown in Table 1, and forward them in the classifier.

Table 1. Features used for each subtask

Subtask	Group	Feature
Ago	Stylometry	3000
Age	Discriminative	5
Gandar	Stylometry	-
Genuer	Discriminative	2

2.3 Classification

We used machine learning tools from the scikit-learn library [5]. For the age and gender subtask we decided, after experimentation with different algorithms, to use a *Support Vector Machine* (SVM) with an RBF kernel and a SVM with a linear kernel respectively. Moreover, we specify that the weights of the SVM kernel should be "balanced" according to the frequency distribution of the samples in the different profiles. This leads to higher weights for the less populated classes, resulting in smaller regularization parameters for these classes and a higher incentive for the SVM to classify it properly.

The best parameters for each SVM classifier were chosen through parameter grid search, using stratified 5-fold cross validation schemes.

We also tried feature scaling and normalization before training the classifier with the concatenated data, for the age subtask, but the corresponding results didn't advocate in favor of this step.

3 Experimental Results

3.1 Dataset and Evaluation

As already mentioned, the core of this year's task is cross-genre identification for three different languages. The dataset distributions for both tasks, showcased for the English language in absolute counts of profiles, are illustrated in Figure 5.



Figure 5. Age and Gender distributions for the English Dataset

In the context of PAN 2016, systems were evaluated using accuracy for both the gender and age classification tasks. An overview of the approaches and results for the author profiling task can be found in [6].

3.2 Results

The evaluation of our system on the two held-out test datasets are shown in Table 2. The best results have been obtained for the English part of the second test set for both subtasks, as highlighted in the table. We can also see that our system is not performing well regarding the age subtask, especially on the Spanish data. Finally, we can see that the results vary greatly per language, e.g. for the Dutch dataset we are consistently subpar on the gender task comparing to the other languages. However, we would need the overall results in order to evaluate our system in comparison with others.

In order to have some kind of comparison, we also evaluated our newly proposed system over the training dataset using a 4-fold cross validation. We pitched our results against our last year's submission which was the 3rd best overall and the best one, on average over all languages, for the gender subtask. The results are showcased in Figure 6. It can be seen that generally the new system outperforms, even marginally, the previous one with the difference being more prominent in the age subtask. Even for the gender subtask, where our previous system excelled in PAN15, we have achieved better results, except for the Dutch dataset.

Dataset	Language	Subtask	Accuracy
Test-1	Dutch	Gender	44.00
	English	Age	30.46
		Gender	53.45
	Spanish	Age	29.69
		Gender	60.94
Test-2	Dutch	Gender	41.60
	English -	Age	55.13
		Gender	69.23
	Spanish -	Age	32.14
		Gender	67.86

 Table 2. Results for the test datasets



Figure 6. Results of the current and last years' system on the current Age(left) and Gender(right) task

4 Conclusions

Although we would need the overall results in order to put ours into perspective, we have gained some initial insights:

- The age subtask is much more difficult than the gender one, as showcased by all evaluations, where both systems consistently perform worse than the gender subtask.
- Regarding the comparison scheme, the results of both systems for the age subtask are far from satisfying with only 50% accuracy. The difference with last year's performance, approximately 66%, may be explained through the nature of the current dataset, as the same methodology achieved only 46% accuracy this year. The volume of the data, the distribution over classes and more importantly the added difficulty of the cross-genre task play an importart role.
- The consistently worse results on the subtasks for the first test dataset, may indicate a genre bias in our model. That is, the data of the 2^{nd} test dataset could be more similar to the train data, than those of the 1^{st} test dataset. This reconfirms the increased complexity of the task when dealing with data from different domains.
- Our system performed better on the English test cases and that is to be expected, as training was mostly done on the English data.Noticing the fluctuations in the accu-

racy results per language maybe a more sophisticated system should have different implementations tailored to the needs of each different language.

4.1 Future Work

These initial results spur us on towards new research work on the Author Profiling task. In the context of our approach we will further evaluate the features used for the age classification subtask specifically. We will utilize the test datasets, when they will be publicly available, in order to find which features deteriorate the performance on the test data. Moreover, using those data we can find which features are important in crossdomain knowledge transfer regarding author profiling traits.

Finally, another approach would be to create target profiles modeling both the age and the gender at the same time, e.g. "male" \land "18-24". This would lead to an increase of the joint accuracy of the two task classifiers and would result in projecting both subtasks in one profile-space of the same nature. That would alleviate the systematic errors imposed through the assumption that age is independent of gender, in terms of linguistic style of texts.

5 Acknowledgments

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

References

- Grivas, A., Krithara, A., Giannakopoulos, G.: Author Profiling Using Stylometric and Structural Feature Groupings—Notebook for PAN at CLEF 2015. In: Cappellato, L., Ferro, N., Jones, G., San Juan, E. (eds.) CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 8-11 September, Toulouse, France. CEUR-WS.org (Sep 2015)
- Li, B., Vogel, C.: Advances in Artificial Intelligence: 23rd Canadian Conference on Artificial Intelligence, Canadian AI 2010, Ottawa, Canada, May 31 – June 2,2010.Proceedings, chap. Improving Multiclass Text Classification with Error-Correcting Output Coding and Sub-class Partitions, pp. 4–15. Springer Berlin Heidelberg, Berlin, Heidelberg (2010)
- Loper, E., Bird, S.: Nltk: The natural language toolkit. In: Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1. pp. 63–70. ETMTNLP '02, Association for Computational Linguistics, Stroudsburg, PA, USA (2002)
- López-Monroy, A., y Gómez, M.M., Escalante, H., Villaseñor-Pineda, L., Villatoro-Tello, E.: INAOE's participation at PAN'13: Author Profiling task—Notebook for PAN at CLEF 2013. In: Forner, P., Navigli, R., Tufis, D. (eds.) CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain (Sep 2013)
- 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)

- Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., Stein, B.: Overview of the 4th Author Profiling Task at PAN 2016: Cross-genre Evaluations. In: Working Notes Papers of the CLEF 2016 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (Sep 2016)
- 7. Rennie, J.D., Shih, L., Teevan, J., Karger, D.: Tackling the poor assumptions of naive bayes text classifiers. In: International Conference on Machine Learning. vol. 20, p. 616 (2003)