Random Forest with Increased Generalization: A Universal Background Approach for Authorship Verification

Notebook for PAN at CLEF 2015

María Leonor Pacheco[1], Kelwin Fernandes[1,2,3], and Aldo Porco[1,4]

[1]Grupo de Inteligencia Artificial, Departamento de Computación y Tecnología de la Información, Universidad Simón Bolívar, Caracas, Venezuela
{07-41302,07-40888,07-41378}@usb.ve
[2]Faculdade de Ciências, Universidade do Porto, Portugal
[3]INESC TEC, Porto, Portugal
kafc@inescporto.pt
[4]Department of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona, Spain
aldo.porco@upf.edu

Abstract This article describes our approach for the Author Identification task introduced in PAN 2015. Given a set of documents written by the same author and a questioned document with an unknown author, the task is to decide whether the questioned document was written by the same author as the other documents or not. Our approach uses Random Forest and a feature-encoding scheme based on the Universal Background Model strategy, building different feature vectors that describe: 1) the complete population of authors in a dataset, 2) the known author, 3) the questioned document and combines the three of them in a single representation.

1 Introduction

Authorship Attribution is the process of attempting to identify the likely authorship of a given document [10]. Important applications of authorship attribution include: plagiarism detection, deducing the writer of inappropriate communications and resolving historical questions of unclear or disputed authorship [7] [5]. A way of approaching authorship attribution is the authorship verification scenario, where we are given a set of documents written by a single author, and we want to determine whether a questioned document is written by the same author or not [9].

The PAN 2015 Author Identification task focuses on the authorship verification problem. In this article we describe our approach for this task, using a feature-encoding scheme inspired on the Universal Background Model and applying Random Forest for prediction.

In section 2 we define the problem of authorship verification formally and introduce relevant notations. Section 3 lists and explains in detail the full set of features considered. Two baseline methods: a simple model based on distances between feature vectors and an implementation of a Gaussian Mixture Model - Universal Background Model are introduced in section 4. Section 5 describes our approach, the feature-encoding scheme and the post-processing done to the Random Forest for probabilistic classification. In Section 6 we present our results and evaluations, both on the train and test corpus. Finally in Section 7 conclusions and future research directions are exposed.

2 Problem Statement

In this section we describe the problem of authorship verification as introduced by the PAN 2015 Author Identification task [9].

Let $P = (D, q)$ be a problem, where D is a small set of documents written by a known author and q is a questioned document whose author we do not know. The Author Identification task consists of determining whether the question document q was written by the same author who wrote the documents on set D or not. In our approach we model this as a probabilistic classification problem, where rather than only outputting the most likely class that the sample should belong to, we obtain a probabilistic output that indicates a degree of certainty between 0.0 and 1.0, corresponding to the probability of a positive answer.

The classification function f is defined as: $f(D, q) = pr$. Where pr is the probability that the questioned document was written by the same author. The size of D will range from 1 to 5 documents.

In this task, we have problems for four different languages: Dutch, English, Greek and Spanish. Evaluations will be measured according to the area under the ROC curve (AUC) of pr and the $c@1$ measure [6]

$$c@1 = \frac{1}{n}(n_c + (\frac{n_u n_c}{n})) \tag{1}$$

Where n refers to the number of problems being evaluated, nc refers to the number of correct answers and nu refers to the number of unanswered problems. For measuring the correctness of an answer, a binary evaluation is performed, where $pr > 0.5$ corresponds to a positive answer, $pr < 0.5$ corresponds to a negative answer and $pr = 0.5$ will be considered as an unanswered problem.

3 Features

This section is devoted to the enumeration and description of the features extracted for this task. We extracted a heterogeneous set of features describing properties related to the style of the author from low level features (e.g. vocabulary diversity, document length, etc) to high level features (e.g. part-of-speech, LDA topics, etc). All of our features are expressed at the author level, considering the total set of documents D for each sample.

3.1 Structure and Extension Features

- **Number of tokens:** minimum, average and maximum number of tokens per document, paragraph and sentence. Also, we extracted the same statistics considering a single occurrence per word, hereafter referred as unique tokens.
- **Number of stop words:** minimum, average and maximum number of stop words (and unique stop words) per document, paragraph and sentence. We considered the stop words dictionaries provided by the Python library many-stop-words.
- **Number of sentences:** minimum, average and maximum number of sentences per paragraph and document.
- **Number of paragraphs:** minimum, average and maximum number of paragraphs per document.
- **Spacing:** minimum, average and maximum number of consecutive spaces, number of consecutive spaces in the beginning/end of the line and number of empty lines.
- **Punctuation:** minimum, average and maximum number of punctuation characters (.,;?¿!¡"’) per document, paragraph and sentence.

3.2 Distributional Features

- **Word distribution:** Frequencies of the words contained in the documents written by the author, divided by the total number of words in them.
- **Character distribution**: Frequencies of the alphanumeric characters in the documents written by the author divided by the total number of characters in his documents. Also, we extracted the minimum, average and maximum number of lowercase characters, uppercase characters and digits per document.
- **Punctuation Bigrams:** Frequency of the punctuation characters bigrams observed in the author documents.

3.3 Linguistic Features

For each author, the following features are extracted independently for each document and then aggregated taking their max, min and average values.

- **Lexical density:** measure of how “dense” is the content, i.e, the ratio between each lexical category (nouns, adjectives, verbs and adverbs) divided by the total number of words.
- **Word diversity:** ratio between the number of lemmas found divided by the total number of words.
- **Lemmas BoW:** frequency of the lemmas.
- **Lemmas diversity:** for each lemma, the number different words mapped to it.
- **Uniqueness:** number of words that appear only one time.
- **Hapax:** number of words that appear only one time and are only used by the current author.

The POS tags and lemmas used were extracted using the Tree Tagger provided by Helmut Schimd [8].

3.4 Topics

– **Word Topics:** Closeness of a document to the K-th LDA topic.
– **Stop word Topics:** Closeness of a document to the K-th LDA topic. Topics are built using only stop words.

4 Baseline

As a way to test our proposal, which will be thoroughly described on section 5, we proposed two baseline models: a simple approach based on distances between feature vectors and an implementation of a Gaussian Mixture Model - Universal Background Model (GMM-UBM) [2], a method commonly used on Speaker Recognition Systems.

4.1 Distance-based approach

In Section 2 we described the Author Identification task as a problem $P = (D, q)$, where D is the set of known documents written by author A, and q is a questioned document.

Let $T = \{P_i(D_i, q_i), ..., P_m(D_m, q_m)\}$ be our complete set of samples in the training set and F_{mxn} the matrix of the complete features extracted for each set of known documents D_i, where m corresponds to the number of problems and and n to the number of features extracted for each D_i. For each row f_j in F, corresponding to the values of feature $j \in \{0, ..., n\}$ for all samples in the training set, we adjust a Gaussian distribution.

We want to determine how *unique* is each author described by $F_{i,j}$ with respect to the total population of samples. For measuring uniqueness, we do: $1.0 - p(author)$.

The lower the probability, the more unique the author described by $F_{i,j}$ and thus the importance of the feature is higher for said author. This is done to derive weights for each feature and normalized per sample so that they sum to 1.0.

For classifying each questioned document, we measure the weighted Euclidean distance between the question document q_i and the set of known documents D_i. An acceptance threshold is trained with all distances, maximizing the classification accuracy.

4.2 Gaussian Mixture Model - Universal Background Model

The Universal Background Model (UBM) is a large Gaussian Mixture Model (GMM) trained to represent the distribution of features for all authors in the dataset. The idea is to derive a model for one specific author by updating the trained parameters in the UBM via a form of Bayesian adaptation [3].

The adaptation is a two-step estimation process, similar to the Expectation Maximization (EM) algorithm. The first step is exactly the same as in the EM algorithm, where estimates of the specific author features are computed for each mixture in the UBM. In the second step of the algorithm the new estimates are combined with the old statistics from the UBM mixture parameters using an adjusted mixing coefficient. This allows mixtures to rely more either on old or new estimates depending on the amount of data from the specific author that they explain [2].

Having trained the UBM and its resulting mixtures on the complete set of known authors, we take the feature vector for a specific author documents $X = \{x1, ..., xt\}$ and compute, for each mixture i in the UBM:

$$Pr(i|x_t) = \frac{wip_i(x_t)}{\sum_{j=1}^{M} w_j p_j(x_t)} \quad (2)$$

We then use $Pr(i|x_t)$ to compute statistics for the weight, mean and variance parameters, following the first step in the EM algorithm:

$$n_i = \sum_{t=1}^{T} Pr(i|x_t) \quad (3)$$

$$E_i(x_t) = \frac{1}{n_i} \sum_{t=1}^{T} Pr(i|x_t) x_t \quad (4)$$

$$E_i(x_t^2) = \frac{1}{n_i} \sum_{t=1}^{T} Pr(i|x_t) x_t^2 \quad (5)$$

Then, these new statistics from the specific author documents are used to update old statistics on the UBM for each mixture i:

$$\widehat{w}_i = [\alpha_i^w n_i / T + (1 - \alpha_i^w) w_i]\gamma \quad (6)$$

$$\widehat{\mu}_i = \alpha_i^m E_i(x) + (1 - \alpha_i^m)\mu_i \quad (7)$$

$$\widehat{\sigma}_i^2 = \alpha_i^v E_i(x^2) + (1 - \alpha_i^v)(\sigma_i^2 + \mu_i^2) - \widehat{\mu}_i^2 \quad (8)$$

The adaptation coefficients are $\{\alpha_i^w, \alpha_i^m, \alpha_i^v\}$ for the weights, means and variance. These are defined by

$$\alpha_i^\rho = \frac{n_i}{n_i + r} \quad (9)$$

Where r is a fixed relevant factor for all parameters ρ, which was set empirically to 16. For tests, we used a fixed set of 2 mixtures

The classification was modeled as a hypothesis test between as proposed by Reynolds et. al [2]. Where: $H1$: The questioned document q belongs to author A and $H0$: The questioned document q does not belong to author A. The decision between these two hypothesis is a likelihood ratio test:

$$\frac{P(q|H0)}{P(q|H1)} \begin{cases} \geq \theta, \text{ accept } H0 \\ < \theta, \text{ reject } H0 \end{cases} \quad (10)$$

5 Random Forest and UBM Decision Strategy

Having as little as one to five document per author, traditional discriminative methods would fail to fit an accurate decision region. Attempting to improve generalization capabilities, we proposed a feature-encoding scheme based on the Universal Background Model (UBM) [2] decision strategy, instead of fitting an entirely new model for each author. Thus, we build a feature vector B for the known set of documents for each language, and a feature vector A for each author. Then, we build a vector U for the questioned document considering it as being written by a new author and encode the problem as:

$$\left\langle \frac{(A_i - U_i)^2 + 1}{(B_i - U_i)^2 + 1} | i \in [0 \dots N) \right\rangle \tag{11}$$

Then, we fed a Random Forest (RF) with each problem. In this way, we are building a model that hierarchically determines the importance of each feature in the identification of authorship. Random Forest is an ensemble discriminative method that trains a set of predictive decision trees to classify a new instance [1]. In contrast to traditional methods for training Decision Trees, RF considers a subset of randomly selected features to train each individual tree.

Each feature would be valued with a number in the interval $[0 \dots 1)$ if the features computed for the unknown document is closer to the author than to the general population, otherwise, it would be valued with a number in $[1 \dots \infty^+)$. This encoding has the advantage that it can model the triad: unknown document - author and population in a single feature vector, making discriminative approaches feasible for this problem (i.e. it does not depend on the number of documents per author but in the number of authors in the dataset). However, as the features grow in an unbounded way with different scales, assuming that all the features lie in the same scale may affect the results. Therefore, we decided to use a Random Forest model which learns the decision region by considering each feature independently [1].

6 Evaluation Results

On the training set for each language, both baseline models and the RF model were scored based on the measure of $AUC * c@1$ using repeated randomly selected subsets: 80% of the samples for training and 20% of the samples for validation. Our RF approach scored higher than both baselines on the four datasets. Resumed details are explained in table 1.

In addition, test sets for each of the languages (Dutch, English, Greek and Spanish) were provided for the competition. We submitted four runs on TIRA [4] for the final evaluation, one for each test set. Table 2 explains our results in detail [9]. According to these results, our approach performed very well on all datasets, reaching an AUC score above 0.75 for all cases and c@1 above 0.5 for three out four tests. We can also observe relatively low run-time on each of the tests performed.

Language	Model	AUC	c@1	AUC * c@1
Dutch	RF	0.935	0.85	0.795
	UBM	0.6	0.6	0.36
	Weighted Distance	0.5	0.5	0.25
English	RF	0.61	0.6	0.365
	UBM	0.55	0.55	0.303
	Weighted Distance	0.5	0.5	0.25
Greek	RF	0.755	0.65	0.491
	UBM	0.65	0.65	0.423
	Weighted Distance	0.7	0.7	0.49
Spanish	RF	1.0	0.95	0.949
	UBM	0.6	0.6	0.36
	Weighted Distance	0.6	0.6	0.36

Table 1. Scoring for proposed model and baselines

Language	Model	AUC	c@1	AUC * c@1	Runtime
Dutch	RF	0.82229	0.75923	0.62431	00:05:08
English	RF	0.76287	0.57429	0.43811	00:15:00
Greek	RF	0.7728	0.6695	0.51739	00:02:01
Spanish	RF	0.9076	0.73	0.66255	00:04:22

Table 2. Performance on the test corpus

7 Conclusion and Future Work

In this work we have presented a supervised learning approach for authorship identification based on Random Forests. Previous attempts to solve this problem focused on the computation and interpretation of distance functions and threshold operations. However, it is a well known problem that nearest neighbor approaches are susceptible to the curse of dimensionality problem. In this sense, as we increase the number of discriminative features in the decision task, the amount of data needed to achieve good results grows exponentially.

Another difficulty related to the proposed problem is the small number of known documents per author, making per-author learning method unfeasible. Therefore, we adopted an discriminative approach with a encoding able to generalize the individual author information by measuring the distance relation between the unknown document, the author's corpora and the entire dataset.

We obtained remarkable results in the Dutch and Spanish tracks, achieving AUC-ROC values of 0.82229 and 0.9076 in the final assessment of the competition, which are the second and third best results obtained in those tracks respectively.

Possible improvements for this approach include studying the separation of documents into paragraphs as a way to introduce more examples, analyze the relevance of the proposed features and include texts from other sources to broaden the dataset, given that our method depends greatly on the number of authors processed.

References

1. Breiman, L.: Random forests. Machine learning 45(1), 5–32 (2001)
2. Douglas A. Reynolds, Thomas F. Quatieri, R.B.D.: Speaker verification using Adapted Gaussian mixture models http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.117.338
3. luc Gauvain, J., hui Lee, C.: Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. IEEE Transactions on Speech and Audio Processing 2, 291–298 (1994)
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. I. N. Bozkurt, O Baglioglu, E.U.: Authorship attribution: Performance of various features and classification methods. In: 22nd international symposium on Computer and information sciences. pp. 1–5. ISCIS 2007, IEEE (2007)
6. Peñas, A., Rodrigo, A.: A simple measure to assess non-response. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1. pp. 1415–1424. HLT ’11, Association for Computational Linguistics, Stroudsburg, PA, USA (2011), http://dl.acm.org/citation.cfm?id=2002472.2002646
7. Sanderson, C., Guenter, S.: Short text authorship attribution via sequence kernels, markov chains and author unmasking: An investigation. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. pp. 482–491. EMNLP ’06, Association for Computational Linguistics, Stroudsburg, PA, USA (2006), http://dl.acm.org/citation.cfm?id=1610075.1610142
8. Schmid, H.: Probabilistic Part-of-Speech Tagging Using Decision Trees http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.28.1139
9. Stamatatos, E., Daelemans, W., Verhoeven, B., Juola, P., Lopez Lopez, A., Potthast, M., Stein, B.: Overview of the Author Identification Task at PAN 2015. In: Working Notes Papers of the CLEF 2015 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (Sep 2015), http://www.clef-initiative.eu/publication/working-notes
10. Zhao, Y., Zobel, J.: Searching with style: Authorship attribution in classic literature. In: Proceedings of the Thirtieth Australasian Conference on Computer Science - Volume 62. pp. 59–68. ACSC ’07, Australian Computer Society, Inc., Darlinghurst, Australia, Australia (2007), http://dl.acm.org/citation.cfm?id=1273749.1273757