# A Machine Learning-based Intrinsic Method for Cross-topic and Cross-genre Authorship Verification

## Notebook for PAN at CLEF 2015

Yunita Sari[1] and Mark Stevenson[2]

NLP Group, Department of Computer Science, University of Sheffield, UK

## 1 Introduction : The task, data set and performance measure

### Author identification task

*"Given a small set (no more than 5, possibly as few as one) of **known** documents by a single person and a **questioned** document, the task is to determine whether the questioned document was written by the same person who wrote the known document set. The genre and/or topic may differ significantly between the known and unknown*

### Data set

The data set consists of author verification problems in four different languages. In each problem, there are some known documents written by single person and only one unknown document. The genre and/or topic between documents may differ significantly.

**Table 1:** Training data set

| Language | Type | Total problems |
|---|---|---|
| Dutch | Cross-genre | 100 |
| English | Cross-topic | 100 |
| Greek | Cross-topic | 100 |
| Spanish | Cross-genre | 100 |

### Performance Measure

The performance of the system will be evaluated using area under the ROC curve (AUC). In addition, the output will also be measured based on c@1 score [4].

$$c@1 = \left(\frac{1}{n}\right) * \left(n_c + \left(n_u * \frac{n_c}{n}\right)\right)$$

Where:

$n$ = number of problems

$n_c$ = number of correct answer

$n_u$ = number of unanswered problems

The overall performance will be evaluated on the product of AUC and c@1

## 2 Methodology : A Machine Learning-based Intrinsic Method

### Textual Representation

Given collection of problems $P = \{P_i : \forall_i \in I\}$ where $I = \{1,2,3,...,n\}$ is the index of $P$. $P_i$ contains exactly one unknown document $K = \{K_j : \forall_j \in J\}$ where $J$ is the index of $K$ and $1 \leq J \leq 5$. Our approach represented each problem $P_i$ as vector $P_i = \{R_1, R_2, ..., R_n\}$ where $n$ is the maximum number of feature types (in our case is six). $R_i$ is the distance of two similar feature vector representation of a set of known documents $K$ and unknown document $U$. If $K$ contains more than one document, then the generated feature vector is an average vector of $J$ documents.

**Table 2:** List of features and comparison measures

| Feature | Model | Comparison method |
|---|---|---|
| (R1) Stylometric features | Average frequency | Min-max similarity |
| (R2) Function words | Ratio of function word to total number of words in the document | Manhattan distance |
| (R3) Character 8-grams | Tf-idf | Cosine similarity |
| (R4) Character 3-grams | Tf-idf | Cosine similarity |
| (R5) Word bigrams | Tf-idf | Cosine similarity |
| (R6) Word unigrams | Tf-idf | Cosine similarity |

### Distance Measure

We experimented with several different comparison measures for computing similarity between a pair of vectors. We noticed particular comparison metrics perform better with certain types of feature, thus we applied different measures for each features type.

### Feature Selection and Classifier

Our authorship identification software was written in Python. We applied feature selection using the Extratreeclassifier and SVM classifiers. The classifier hyperparameters were optimized using GridSearchCV. Scikit-learn library was used for both feature selection and classification.

### Stylometric Features

Along with six Stylometry features, we also implemented three readability measures including: Flesch-Kincaid Reading Ease [3], Flesch-Kincaid Grade Level [3], Gunning-Fog Index [2].

## 3 Evaluation on Training Corpora

- The approach was evaluated on the training data using 10–fold cross validation.
- We did not perform the verification task on Greek data due to character set incompatibility issues.
- The best result was achieved on the Spanish data set which has more known documents compared to other sub-language corpora.

**Table 3:** 10-fold cross validation on the training corpora

| Data set | AUC | C@1 | finalScore |
|---|---|---|---|
| English | 0.662 | 0.606 | 0.401 |
| Dutch | 0.618 | 0.553 | 0.342 |
| Spanish | 0.846 | 0.807 | 0.683 |

## 4 Evaluation on Testing Corpora

**Table 4:** Result on test data set

| Data set | AUC | C@1 | finalScore | Runtime |
|---|---|---|---|---|
| English | 0.4011 | 0.5 | 0.20055 | 00:05:46 |
| Dutch | 0.61306 | 0.62075 | 0.38056 | 00:02:03 |
| Spanish | 0.7238 | 0.67 | 0.48495 | 00:03:47 |

- The verification system performed best on the Spanish data set
- The performance of supervised learning-based system heavily relies on the amount of training data.
- In terms of runtime, our approach was generally efficient since all necessary processing steps were performed in the training phase.

## 5 Conclusion

- Limited number of known documents per problem and significant differences in genre/topic caused the verification task is
- English data set was derived from Project Gutenberg's opera play scripts which are an unusual type of text.
- The most challenging part of this task was to find suitable features which could capture the differences between documents.
- Applying feature selection were beneficial and greatly improved the accuracy of the classifier.

## References

**1.** PAN Authorship Identification Task 2015, https://www.uniweimar.de/medien/webis/events/pan-15/pan15-web/author-identification.html

**2.** Gunning, R.: The Technique of Clear Writing. McGraw-Hill (1952)

**3.** Kincaid, J.P., Fishburne, R.P., Rogers, R.L., Chissom, B.S.: Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Tech. Rep. February (1975)

**4.** 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. pp. 1415–1424 (2011)

**URL:** http://nlp.shef.ac.uk **e-mail:**{y.sari[1], mark.stevenson[2]}@sheffield.ac.uk