

An Author Verification Approach Based on Differential Features

Alberto Bartoli Alex Dagri Andrea De Lorenzo
Eric Medvet* Fabiano Tarlao

Department of Engineering and Architecture
University of Trieste
Italy



September 9th, 2015

Table of Contents

1 Problem statement

2 Our approach

3 Analysis



Author verification

Input: problem instance $\langle K, u, L \rangle$

- K is a set of documents authored by the same author
- u is a document whose author is unknown
- L is the language of $K \cup \{u\}$



Author verification

Input: problem instance $\langle K, u, L \rangle$

- K is a set of documents authored by the same author
- u is a document whose author is unknown
- L is the language of $K \cup \{u\}$

Goal: determine (with confidence in $[0, 1]$) if u is authored by the same author of K

- 0 means “sure that no”
- 1 means “sure that yes”
- 0.5 means “don’t know”



Author verification

Input: problem instance $\langle K, u, L \rangle$

- K is a set of documents authored by the same author
- u is a document whose author is unknown
- L is the language of $K \cup \{u\}$

Goal: determine (with confidence in $[0, 1]$) if u is authored by the same author of K

- 0 means “sure that no”
- 1 means “sure that yes”
- 0.5 means “don’t know”

A *training set* is available of solved problem instances (each solution is either 0 or 1)



Nature of the data

- 4 languages: English, Dutch, Greek, Spanish
- different topics in documents
- different sizes of K



Assessment

Using a *testing set* of solved problem instances unavailable to PAN2015 participants, by means of:

- area under the curve (AUC)
- $c@1$, takes into account “don't know” answers



Table of Contents

- 1 Problem statement
- 2 Our approach
- 3 Analysis



In a nutshell

- 9 *groups* of homogeneous numerical features, each feature f to be extracted from a document
- actual feature value is the *difference* between average value on known docs in K and on unknown doc u , i.e., feature F is extracted from a problem instance
- a regressor applied to features of each group
- regressor outcomes are averaged (*ensemble* of regressors)



In a nutshell

- 9 *groups* of homogeneous numerical features, each feature f to be extracted from a document
- actual feature value is the *difference* between average value on known docs in K and on unknown doc u , i.e., feature F is extracted from a problem instance
- a regressor applied to features of each group
- regressor outcomes are averaged (*ensemble* of regressors)

Each language considered in isolation



Feature groups

- 1 Word *n*grams (WG)
- 2 Character *n*grams (CG)
- 3 POS (part-of-speech) tag *n*grams (PG)
- 4 Word lengths (WL)
- 5 Sentence lengths (SL)
- 6 Sentence length *n*grams (SG)
- 7 Word richness (WR) (one single feature)
- 8 Punctuation *n*grams (MG)
- 9 Text shape *n*grams (TG)



Feature groups

- 1 Word *n*grams (WG)
- 2 Character *n*grams (CG)
- 3 POS (part-of-speech) tag *n*grams (PG)
- 4 Word lengths (WL)
- 5 Sentence lengths (SL)
- 6 Sentence length *n*grams (SG)
- 7 Word richness (WR) (one single feature)
- 8 Punctuation *n*grams (MG)
- 9 Text shape *n*grams (TG)



Sentence lengths (SL)

Idea: capture the mixture of sentence lengths used by the author

- sentences obtained by splitting by . ; : ? !
- feature values are number of n -word long sentences, with $1 \leq n \leq 40$



Sentence lengths (SL)

Idea: capture the mixture of sentence lengths used by the author

- sentences obtained by splitting by . ; : ? !
- feature values are number of n -word long sentences, with $1 \leq n \leq 40$

Here I am. And I feel very good. Yes, very very good.
But it's time to end this slide description.

n	1	2	3	4	5	...	9	...	40
# of	0	0	1	1	1	...	1	...	0



Sentence length *n*grams (SG)

Idea: capture the rhythm in terms of sentence length

- 4 sentence length intervals, according to quartiles on the training set
- a symbol represents an interval (Tiny, Short, Medium, Long)
- feature values are occurrences of bigrams of symbols ($n \leq 2$)



Sentence length n grams (SG)

Idea: capture the rhythm in terms of sentence length

- 4 sentence length intervals, according to quartiles on the training set
- a symbol represents an interval (Tiny, Short, Medium, Long)
- feature values are occurrences of bigrams of symbols ($n \leq 2$)

Here I am. And I feel very good. Yes, very very good.
But it's time to end this slide description.

↓
SSSM

bigram	T	S	M	...	SS	SM	MM	...
# of	0	3	1	...	2	1	0	...



Text shape *n*grams (TG)

Idea: capture usage of digits, upper- and lowercase

- sequences of digits \rightarrow symbol n
- sequences of lowercase alphabetic \rightarrow symbol l
- sequences of lowercase alphabetic with first uppercase \rightarrow symbol u
- sequences of lowercase alphabetic with ≥ 2 uppercase \rightarrow symbol w
- feature values are occurrences of trigrams of symbols ($n \leq 3$)



Text shape *n*grams (TG)

Idea: capture usage of digits, upper- and lowercase

- sequences of digits \rightarrow symbol *n*
- sequences of lowercase alphabetic \rightarrow symbol *l*
- sequences of lowercase alphabetic with first uppercase \rightarrow symbol *u*
- sequences of lowercase alphabetic with ≥ 2 uppercase \rightarrow symbol *w*
- feature values are occurrences of trigrams of symbols ($n \leq 3$)

Here I am. And I feel very good. Yes, very very good.
But it's time to end this slide description.

↓

uwluwllllullllulllllllll

↓

trigram	u	w	l	...	ull	...
# of	4	2	15	...	2	...

Feature selection

- only for features of largest groups (WG, CG, PG, and TG—word, chars, punct., text-shape grams)
- using all the documents of the training set (of a given language)
- n_{sel} features with greatest average values are selected ($n_{\text{sel}} = 100$ for TG, $n_{\text{sel}} = 500$ for the others)



Feature selection

- only for features of largest groups (WG, CG, PG, and TG—word, chars, punct., text-shape grams)
- using all the documents of the training set (of a given language)
- n_{sel} features with **greatest average values** are selected ($n_{\text{sel}} = 100$ for TG, $n_{\text{sel}} = 500$ for the others)

Aim: avoid overfitting on unusual features (e.g., a word which appears on just one doc)



Feature normalization

Feature values are normalized within group G for the input document d :

$$f_i(d) := \frac{f_i(d)}{\sum_{f_j \in G} f_j(d)}$$



Actual feature value: difference

Actual feature value $F_i(\cdot)$ is the absolute difference between average value on known docs in K and value on unknown doc u :

$$F_i(\langle K, u, L \rangle) = \text{abs} \left(\frac{\sum_{k \in K} f_i(k)}{|K|} - f_i(u) \right)$$



Actual feature value: difference

Actual feature value $F_i(\cdot)$ is the **absolute difference** between average value on known docs in K and value on unknown doc u :

$$F_i(\langle K, u, L \rangle) = \text{abs} \left(\frac{\sum_{k \in K} f_i(k)}{|K|} - f_i(u) \right)$$



Actual feature value: difference

Actual feature value $F_i(\cdot)$ is the absolute difference between **average value on known docs in K** and value on unknown doc u :

$$F_i(\langle K, u, L \rangle) = \text{abs}\left(\frac{\sum_{k \in K} f_i(k)}{|K|} - f_i(u)\right)$$



Actual feature value: difference

Actual feature value $F_i(\cdot)$ is the absolute difference between average value on known docs in K and **value on unknown doc u** :

$$F_i(\langle K, u, L \rangle) = \text{abs}\left(\frac{\sum_{k \in K} f_i(k)}{|K|} - f_i(u)\right)$$



Actual feature value: difference

Actual feature value $F_i(\cdot)$ is the absolute difference between average value on known docs in K and value on unknown doc u :

$$F_i(\langle K, u, L \rangle) = \text{abs}\left(\frac{\sum_{k \in K} f_i(k)}{|K|} - f_i(u)\right)$$

Aim: capturing the difference in writing style between known and unknown documents



Variant

Weight by feature value on unknown document:

$$F'_i(\langle K, u, L \rangle) = \frac{F_i(\langle K, u, L \rangle)}{f_i(u)} \quad (1)$$



Regressors

Explored 3 options:

- trees
- random forest
- SVM

Recall: a regressor is applied to a single feature group, outcome is then averaged across groups



Table of Contents

- 1 Problem statement
- 2 Our approach
- 3 Analysis



Variant choice

6 approach variants:

- 2 kinds of feature ($F_i(\cdot)$ and $F'_i(\cdot)$)
- 3 regressors



Variant choice

6 approach variants:

- 2 kinds of feature ($F_i(\cdot)$ and $F'_i(\cdot)$)
- 3 regressors

Comparison on the training set (partitioned by language) with a *leave-one-out* procedure



Variant choice: results on the training set

Variant	$c@1$				AUC			
	EN	DU	GR	SP	EN	DU	GR	SP
RF- F	0.67	0.74	0.77	0.94	0.718	0.707	0.808	0.992
RF- F'	0.58	0.66	0.77	0.95	0.584	0.776	0.796	0.989
SVM- F	0.48	0.67	0.69	0.92	0.513	0.707	0.754	0.978
SVM- F'	0.45	0.62	0.66	0.86	0.584	0.645	0.732	0.936
Tree- F	0.69	0.70	0.53	0.94	0.725	0.708	0.557	0.951
Tree- F'	0.56	0.62	0.69	0.97	0.526	0.595	0.699	0.992



Variant choice: results on the training set

Variant	$c@1$				AUC			
	EN	DU	GR	SP	EN	DU	GR	SP
RF- F	0.67	0.74	0.77	0.94	0.718	0.707	0.808	0.992
RF- F'	0.58	0.66	0.77	0.95	0.584	0.776	0.796	0.989
SVM- F	0.48	0.67	0.69	0.92	0.513	0.707	0.754	0.978
SVM- F'	0.45	0.62	0.66	0.86	0.584	0.645	0.732	0.936
Tree- F	0.69	0.70	0.53	0.94	0.725	0.708	0.557	0.951
Tree- F'	0.56	0.62	0.69	0.97	0.526	0.595	0.699	0.992

- RF- F is, in general, the best choice
- English is hard!



Additional analysis

Goals:

- which are best feature groups?
- are there good groups of groups?



Additional analysis

Goals:

- which are best feature groups?
- are there good groups of groups?

Ultimate goal:

- is there a better combination for English?



Additional analysis: English

Is there a better combination for English?



Additional analysis: English

Is there a better combination for English?

Eventually, we made up an answer: “yes, just for English, we’ll use RF- F' with *only 3 features groups* (MG+CG+SL)” (punctuation, chars, sentence length)



Additional analysis: English

Is there a better combination for English?

Eventually, we made up an answer: “yes, just for English, we’ll use RF- F' with *only 3 features groups* (MG+CG+SL)” (punctuation, chars, sentence length)

Results on testing set

Method	Language	c@1	AUC	Score	Ranking
RF- F on MG+CG+SL	EN	0.56	0.578	0.323	10/18
RF- F on all groups	DU	0.69	0.751	0.518	4/17
RF- F on all groups	GR	0.66	0.698	0.459	7/14
RF- F on all groups	SP	0.83	0.932	0.773	1/17



Thanks!

