Authorship Verification for Different Languages, Genres and Topics

Oren Halvani, Christian Winter, Anika Pflug

Fraunhofer Institute for Secure Information Technology (SIT), Darmstadt, Germany
Department of Computer Science Technische Universität Darmstadt, Germany

Fraunhofer

SIT

Bundesministerium
für Bildung
und Forschung

CASED

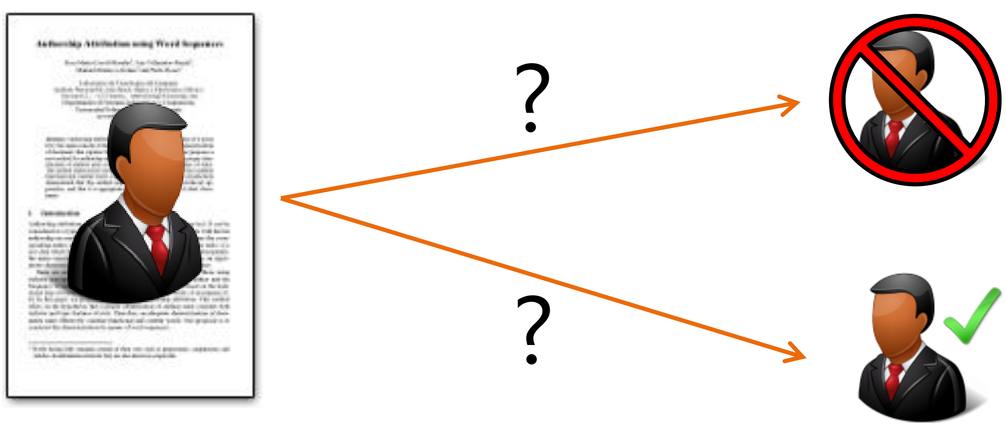
OVERVIEW

- Motivation
- Features
- Corpora
- Our AV method
- Evaluation
- Observations / benefits / future work

MOTIVATION

- Authorship Verification (**AV**) is an important sub discipline of digital text forensics

- Task of AV: Decide if a questionable document was truly written by the stated author, or not…



MOTIVATION

- AV has many application scenarios…
  - **Detect commercial fraud** (such as fictive insurance claims invented by a field agent of an insurance company)
  - **Multiple account detection / User verification** (e.g. WhatsApp, Skype, Facebook, etc.)
  - **Leakage prevention** (e.g. detect if employees leak confidential information through unapproved communication channels)

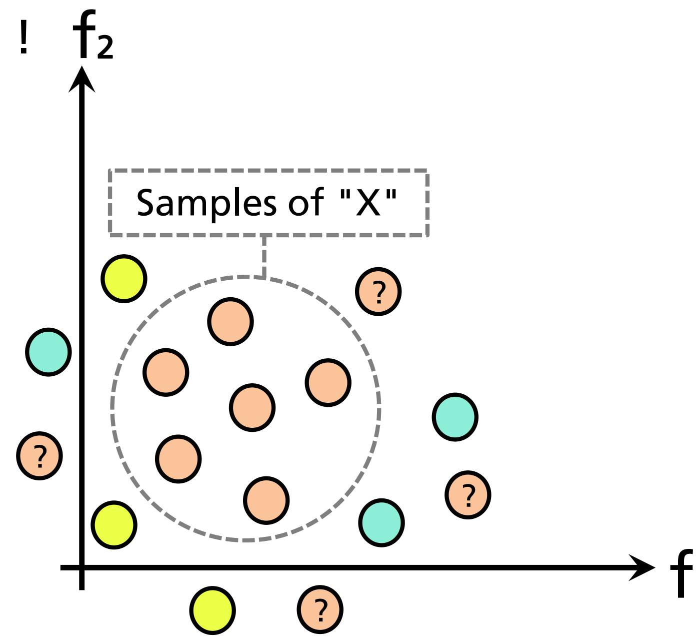
MOTIVATION

- However, AV is also a very challenging task !

- Imagine we have six sample documents of an author "X"…

- **Problem 1:** There might be many other documents of "X", which we don’t have

- **Problem 2:** There are billions of other authors who can claim they are "X"



- **Problem 1 + 2 :** How can we accept unseen documents of "X" and simultaneously reject those of other authors?

FEATURES

- The writing style of an author is individual…
  …or conversely: Writing style cannot be formalized !

- Therefore, heuristics are needed in order to perform AV

- One heuristic (perhaps the only possible one) is to use a set of style markers (features) which aim to model the writing style of an author

FEATURES

- We use only text-surface features…

Example: Halvani $\xrightarrow{n = 3}$ (Hal, alv, lva, …)

Feature Category	Parameters
F_1: Punctuation n-grams	$n \in \{1, 2, \ldots, 10\}$
F_2: Character n-grams	$n \in \{1, 2, \ldots, 10\}$
F_3: $n\%$ frequent tokens	$n \in \{5, 10, \ldots, 50\}$
F_4: Token k-prefixes	$k \in \{1, 2, 3, 4\}$
F_5: Token k-suffixes	$k \in \{1, 2, 3, 4\}$
F_6: Token k-prefix n-grams	$n \in \{2, 3, 4\}$, $k \in \{1, 2, 3, 4\}$
F_7: Token k-suffix n-grams	$n \in \{2, 3, 4\}$, $k \in \{1, 2, 3, 4\}$
F_8: n-prefixes–k-suffixes	$n, k \in \{1, 2, 3, 4\}$
F_9: n-suffixes–k-prefixes	$n, k \in \{1, 2, 3, 4\}$

CORPORA

- In our scheme we consider various corpora (annotated document collections), extend over different languages, genres and topics

- We compiled corpora from different **online** sources (forums, news portals, social networks, etc.) as well as **offline** sources (e-Mails, degree theses, magazine articles, etc.)

CORPORA

- In the learning phase of our AV method we treat all corpora of one language as a single corpus such that each language represents a training corpus…

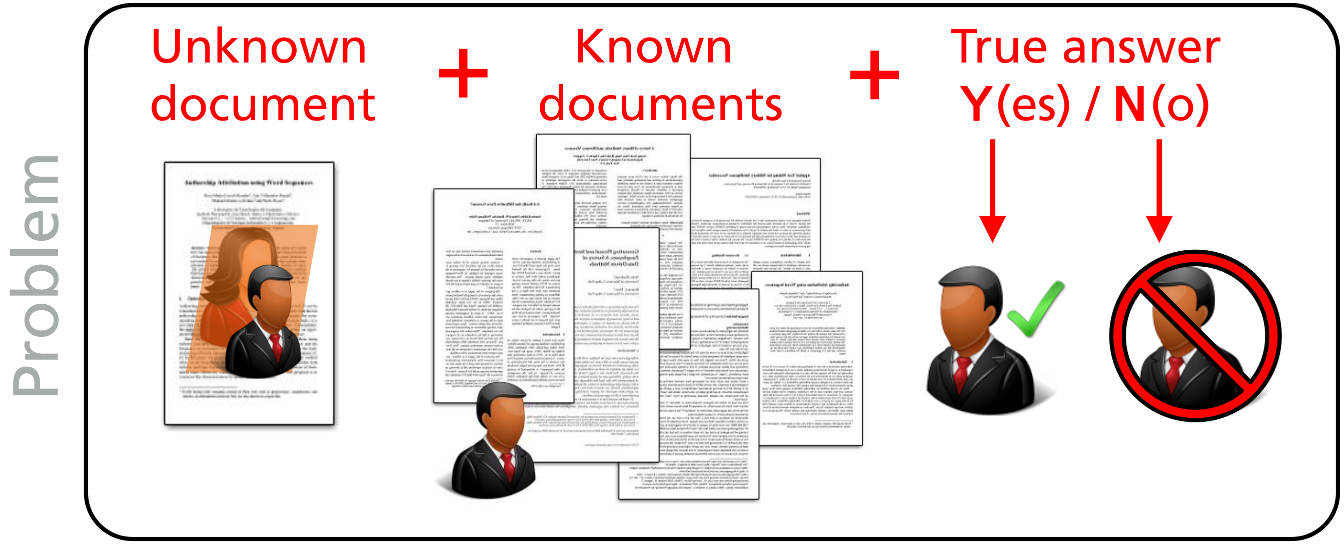
1. Dutch (NL)
2. English (EN)
3. Greek (GR)
4. Spanish (SP)
5. German (DE)

- …this helps to generalize across different genres and topics

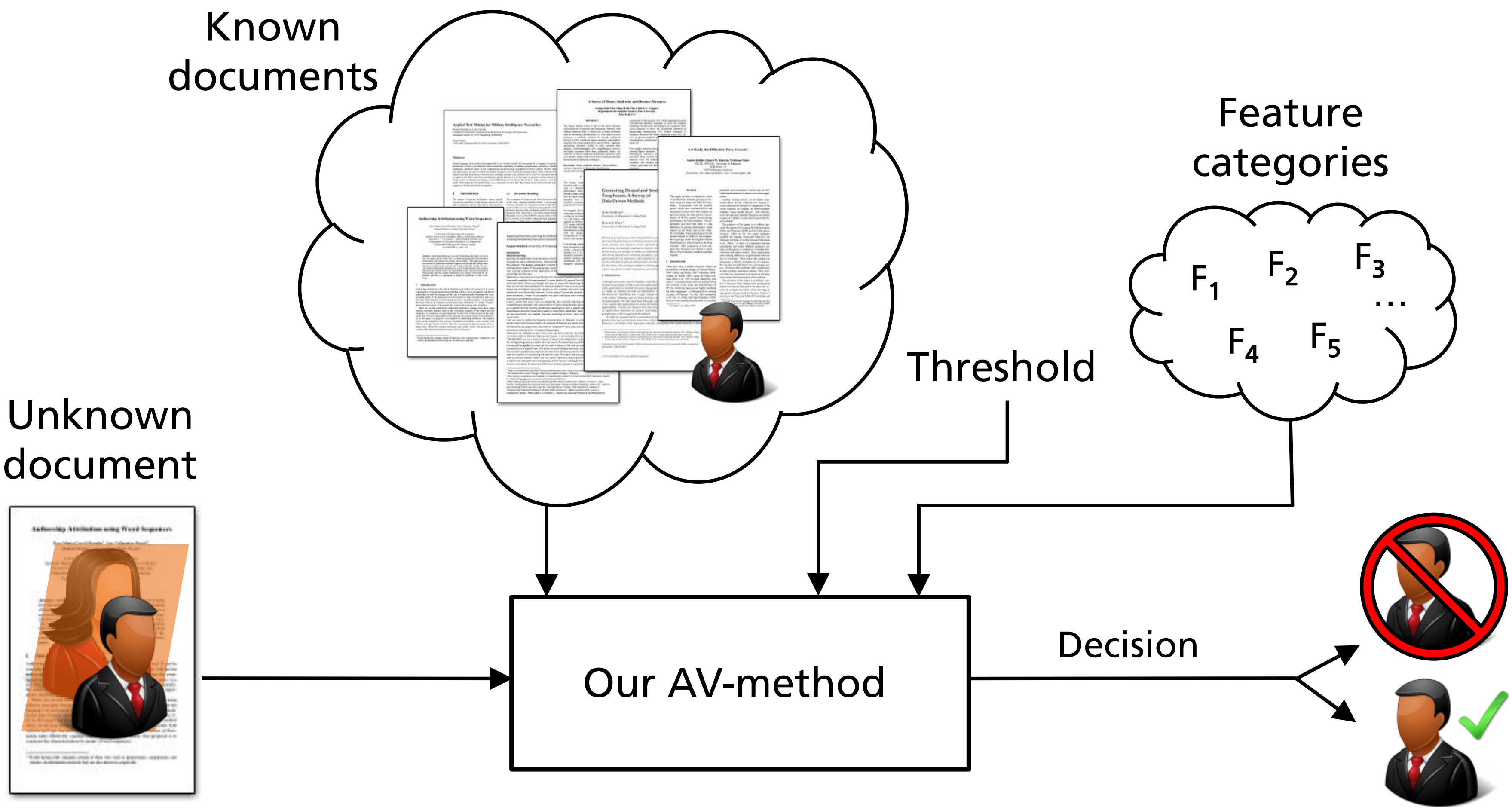
CORPORA

- All corpora follow a unique format, where each corpus comprises **n** so-called "problems"

- A problem consists of an unknown document, a set of known documents and the true answer regarding the questioned authorship…



OUR AV METHOD



OUR AV METHOD

Learning phase (training corpora, feature categories & parameters)

foreach(training corpus = language)
{
Model$_1$ (optimal configurations = parameters & threshold)

Model$_2$ (optimal ensemble = combination of feature categories)
}

Testing (problem, Model$_1$, Model$_2$)

1.) Construct feature vectors and calculate similarity scores

2.) Classify problem as **Y** or **N**

OUR AV METHOD: LEARNING PHASE

Model$_1$ = ( )

foreach(feature category) {
  foreach(feature category parameter) {
    Scores = foreach(problem) { Construct feat. vectors, calculate sim. scores }
    Determine EER-Threshold(Scores)
    Predictions = foreach(problem) { $\text{classify}(\rho) = \begin{cases} \texttt{Y} & \text{if } s_\rho > \text{EER-threshold} \\ \texttt{N} & \text{otherwise} \end{cases}$ }

    Accuracy = $\dfrac{\#(\text{correct answers})}{\#(\text{all problems in the corpus})}$
  }
  Model$_1$.Update(accuracy)
}

return **Model**$_1$ = Optimal configurations = parameters & threshold

OUR AV METHOD: LEARNING PHASE

- **For each problem:** Construct feature vectors, calculate similarity scores

Unknown document

$$\mathbf{X} = (x_1, x_2, x_3, \dots, x_n)$$

$$x_1, y_1 = \frac{\text{Frequency of all "The"}}{\text{Number of all tokens}}$$

Concatenate

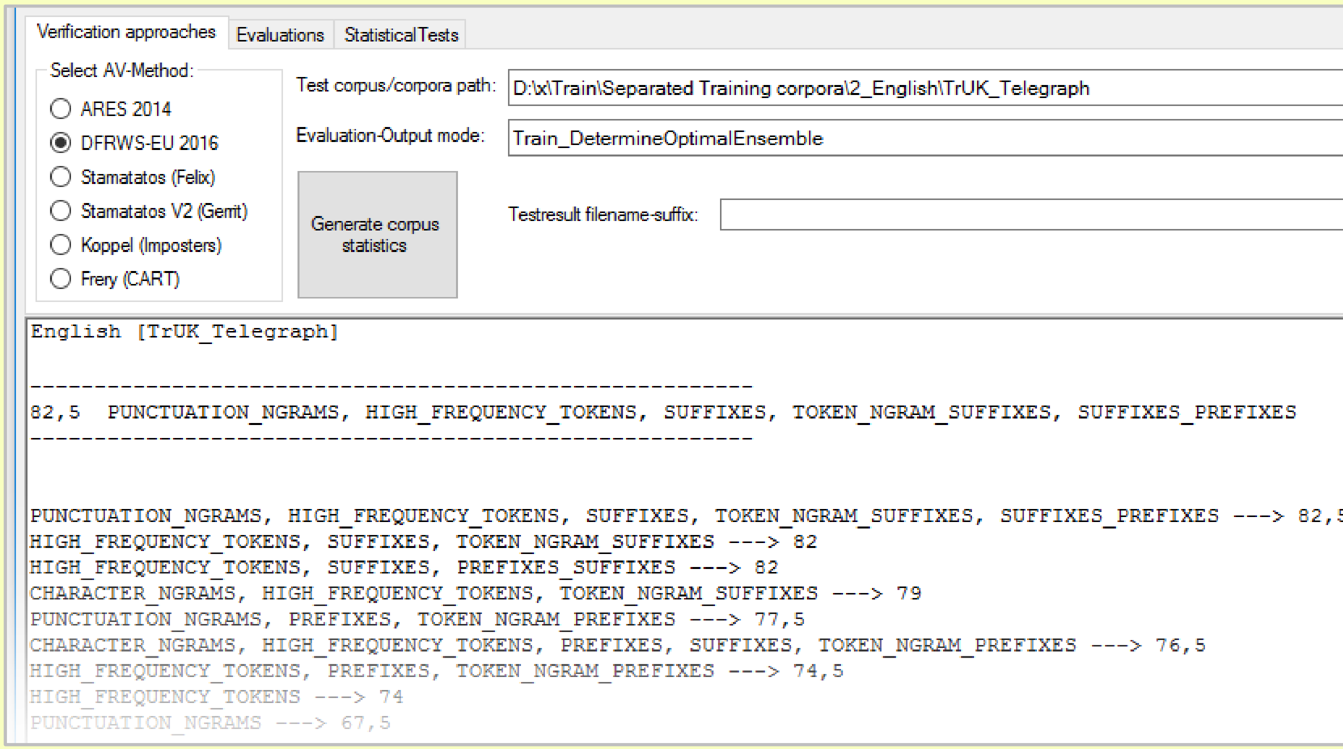
Known document

$$\mathbf{Y} = (y_1, y_2, y_3, \dots, y_n)$$

$$s_\rho = \mathrm{sim}(X, Y) = \frac{1}{1 + \sum_{i=1}^{n} |x_j - y_j|}$$

OUR AV METHOD: LEARNING PHASE

Model$_2$ = ( )



Calculate all possible ensembles…

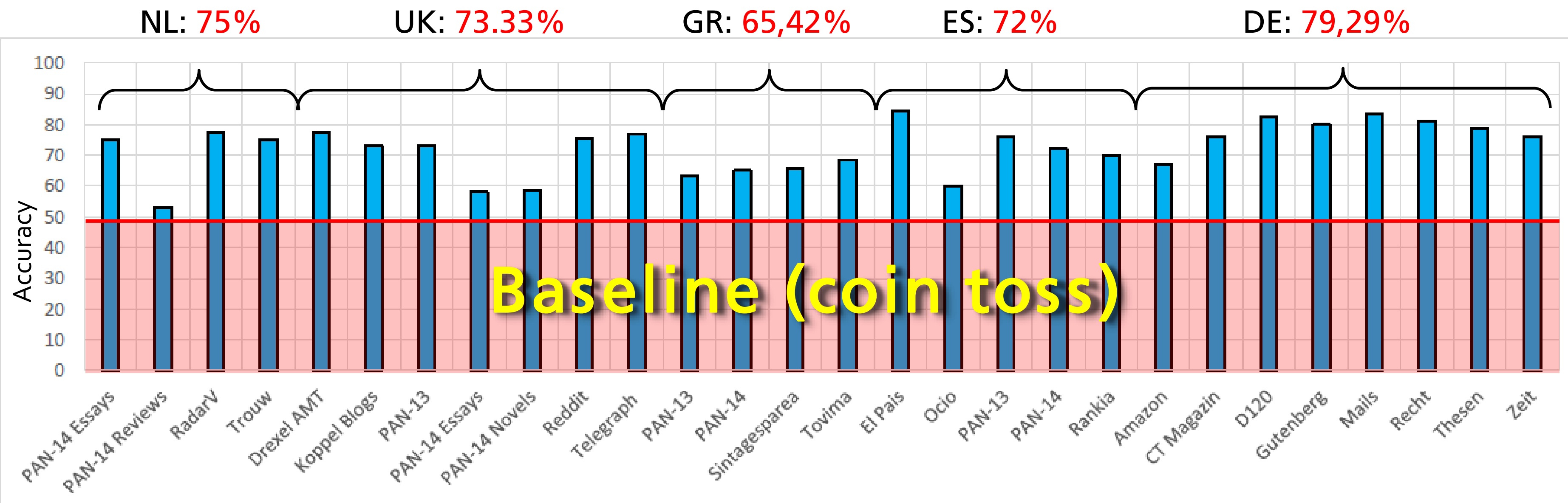
return **Model$_2$** = Optimal ensemble = combination of feature categories

EVALUATION

- We evaluated our method on 28 test corpora (4,525 problems, distributed over 5 languages, 16 genres and > 1000 mixed topics
- **Internal evaluation:** Our method against 2 other promising AV methods of Erwan Moreau and Efstathios Stamatatos. → Both evaluated their AV methods on our corpora
- **External evaluation:** Our method against 17 participants and 4 baselines within an international AV competition (PAN.Webis.de)

EVALUATION (INTERNAL)

- Results of the test set evaluation regarding the 28 test corpora:



- Overall median accuracy:
75% (our approach), 70% (Moreau), 69.3% (Stamatatos)

EVALUATION (EXTERNAL)

- Our AV method was also evaluated at the PAN 2015 competition…

☞ PAN.Webis.de

PAN 2015

This is the 13th evaluation lab on uncovering plagiarism, authorship, and social software misuse. PAN will be held as part of the CLEF conference in Toulouse, France, on September 8-11, 2015. Evaluations will commence from January till June. We invite you to take part in any of the three tasks shown below.

Learn more » Register now » 158 already signed up

Plagiarism Detection

Given a document, is it an original?

This task is divided into **source retrieval** and **text alignment**. Source retrieval is about searching for likely sources of a suspicious document. Text alignment is about matching passages of reused text between a pair of documents.

Author Identification

Given a document, who wrote it?

This task focuses on **authorship verification** and methods to answer the question whether two given documents have the same author or no. This question accurately emulates the real-world problem that most forensic linguists face every day.

Author Profiling

Given a document, what're its author's traits?

meaning cloud

Sponsor

This task is concerned with predicting an author's demographics from her writing. For example, an author's style may reveal her **age, gender, and personality**.

EVALUATION (PAN 2015)

- Results of the PAN 2015 competition (evaluation on 1,265 problems)
- **Note:** Performance measure is the product of AUC and C@1 (known measure in the AV field)
- **Observation:** Our AV method is robust in terms of languages, compared to majority of all approaches

Source: PAN15-AI-Overview Paper

		Language				
Rank	Team	NL	EN	GR	SP	Average
1	Bagnall	0,451	0,614	0,75	0,721	0,628
2	Moreau et al.	0,635	0,453	0,693	0,661	0,606
3	Pacheco et al.	0,624	0,438	0,517	0,663	0,558
4	Huerlimann et al.	0,616	0,412	0,599	0,539	0,538
–	PAN15-ENSEMBLE	0,426	0,468	0,537	0,715	0,532
5	Bartoli et al.	0,518	0,323	0,458	0,773	0,506
6	Gutierrez et al.	0,329	0,513	0,581	0,509	0,478
7	Halvani et al.	0,455	0,458	0,493	0,441	0,462
8	Kocher & Savoy	0,218	0,508	0,631	0,366	0,416
–	PAN14-BASELINE-2	0,191	0,409	0,412	0,683	0,405
9	Maitra et al.	0,518	0,347	0,357	0,352	0,391
10	Castro-Castro et al.	0,247	0,52	0,391	0,329	0,365
–	PAN13-BASELINE	0,242	0,404	0,384	0,367	0,347
11	Gomez-Adorno et al.	0,39	0,281	0,348	0,281	0,323
–	PAN14-BASELINE-1	0,255	0,249	0,198	0,443	0,28
12	Sari & Stevenson	0,381	0,201	-	0,485	0,25
13	Pimas et al.	0,262	0,257	0,23	0,24	0,247
14	Solorzano et al.	0,153	0,259	0,33	0,218	0,235
15	Posadas-Duran et al.	0,132	0,4	-	0,462	0,226
16	Nikolov et al.	0,089	0,258	0,454	0,095	0,201
17	Vartapetiance & Gillam	0,262	-	0,212	0,348	0,201
18	Mechti et al.	-	0,247	-	-	0,063

OBSERVATIONS

- AV works well with ~5KByte (noise-free) texts

- In general we observed:
  \+ News articles, e-Mails, forum postings
  – Essays, novels

- Character n-grams seem to be the most powerful features

  → However, these features are not independent of the topic of the text and thus, should be reconsidered !

BENEFITS

Our AV method provides a number of benefits:

- **Universal:** Applicable for many Indo-European languages such as English, German, Spanish, Greek, Dutch (also French, Polish and Swedish)
- **Independent:** Doesn't make use of linguistic resources such as wordlists, ontologies, thesauruses, language models, etc.
- **Low runtime:** Simple & fast algorithm (no machine learning or deep linguistic processing)

→ Verification runtime of a problem = near real-time !

FUTURE WORK

- Discard features that potentially carry semantic information…

- Try to locate the writing style in a more comprehensible manner

  → This will help to establish the AV Method at court

- Investigate the robustness of our AV method against text modifications such as insertion / deletion of words, paraphrasing…

Thank you for listening ;-)


Fraunhofer
SIT

CASED

M.Sc. Inf.
Oren Halvani

Media Security and IT Forensics
Fraunhofer Institute for Secure Information Technology SIT

Rheinstrasse 75 | 64295 Darmstadt | Germany
Phone +49 6151 869-211 | Fax +49 6151 869-224
Mobile +49 179 2838686 | oren.halvani@sit.fraunhofer.de
www.sit.fraunhofer.de

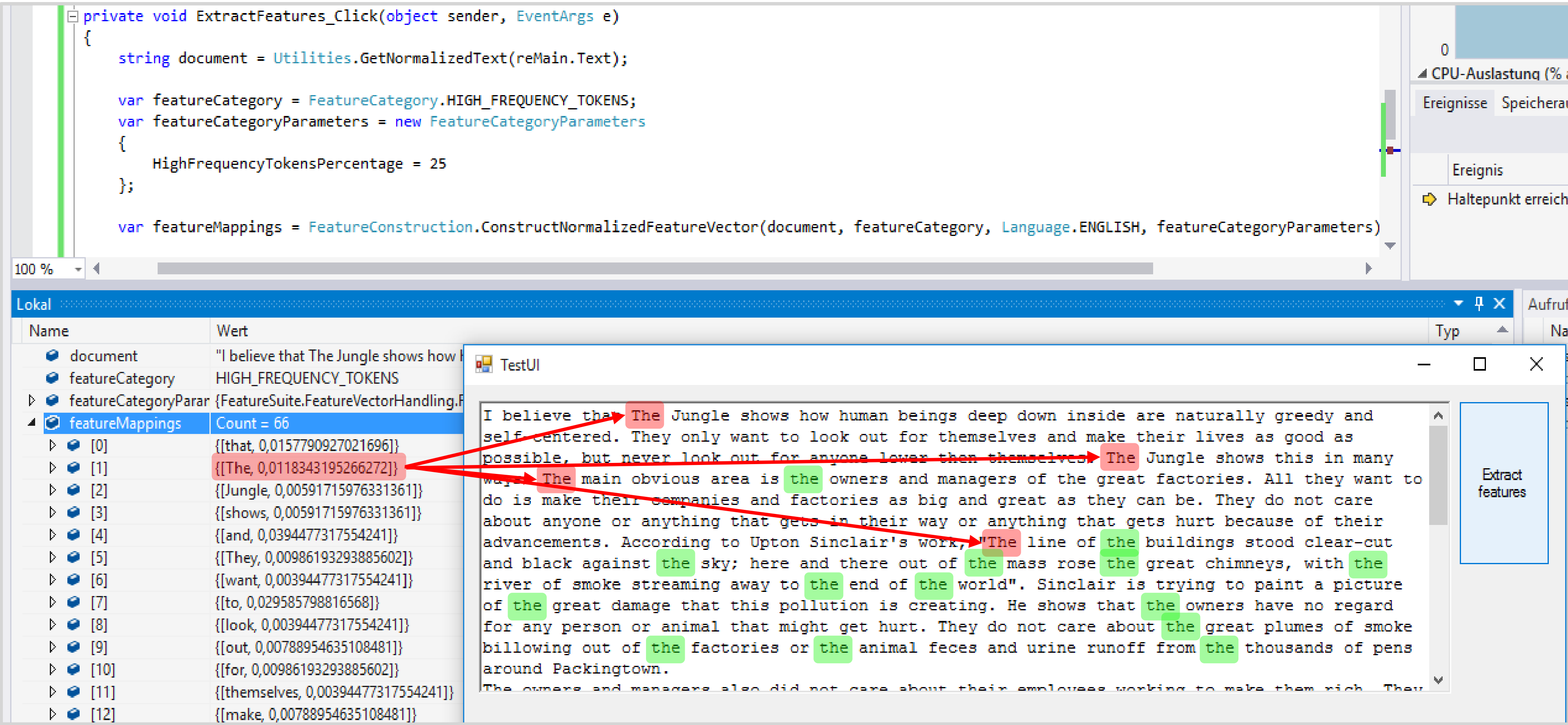
BACKUP SLIDES: PREPROCESSING

- Before applying our AV method on a problem, all involved documents undergo **noise reduction** and **normalization**

Remove tags (HTML, CSS, etc.), tokens consisting of a mix of symbols and only few printable letters (e.g. tbl:XY-19!) as well as digits. Reason: they don't carry any stylistic information of authors

Substitute non-printing control characters (newlines, tabs, etc.) as well as successive blanks by one blank. Furthermore, equalize lengths of all training documents.

BACKUP SLIDES: FEATURE EXTRACTION



BACKUP SLIDES: PAN 2015 CORPUS STRUCTURE

PAN-2015 Corpus

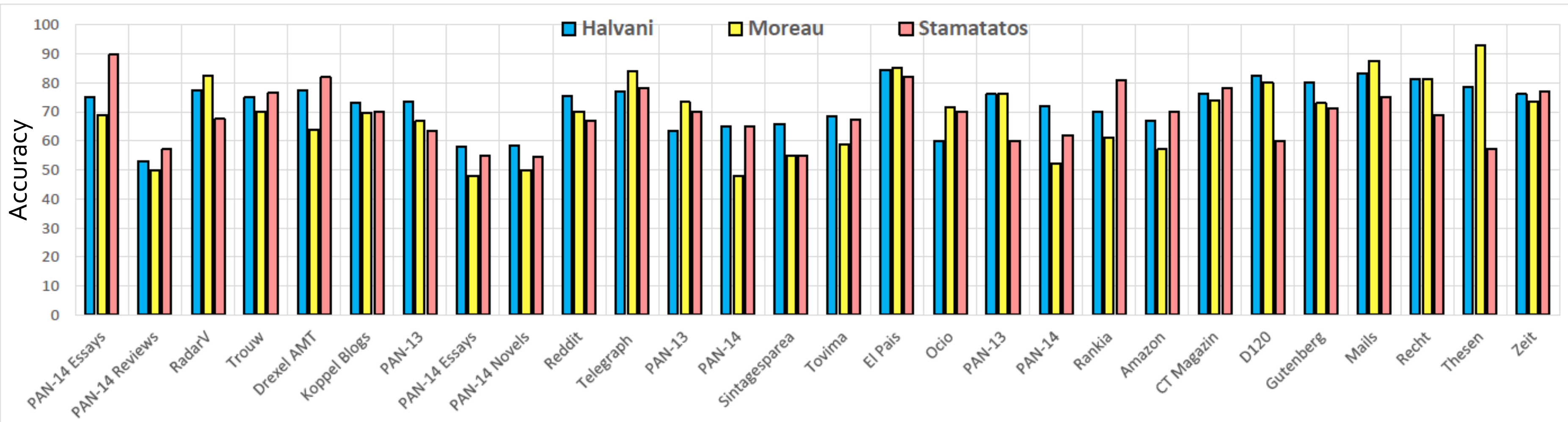
	Language	Type	#Problems	#Docs	Avg. known docs per problem	Avg. words per document
Training	Dutch	cross-genre	100	276	1.76	354
	English	cross-topic	100	200	1.00	366
	Greek	cross-topic	100	393	2.93	678
	Spanish	mixed	100	500	4.00	954
Evaluation	Dutch	cross-genre	165	452	1.74	360
	English	cross-topic	500	1000	1.00	536
	Greek	cross-topic	100	380	2.80	756
	Spanish	mixed	100	500	4.00	946
TOTAL			**1265**	**3701**	**1.93**	**641**

All corpora are balanced (positive/negative problems)

Source: PAN15-AI-Overview Slides

BACKUP SLIDES: EVALUATION (INTERNAL)

- Results of the test set evaluation regarding the 28 test corpora:



- **Outperformed cases:** 19 / 28 (Halvani vs. Moreau), 14 / 28 (Halvani vs. Stamatatos), 10 / 28 (Halvani vs. Moreau & Stamatatos)