Sprachunabhängige Autorschafts-Verifikation

Oren Halvani, Martin Steinebach, Ralf Zimmermann

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

CASED

Fraunhofer SIT

© Fraunhofer

ÜBERBLICK

- Motivation
- Einführung: Autorschafts-Verifikation
- Feature-Kategorien
- Unser Verfahren
- Evaluierung
- Zusammenfassung / Ausblick

MOTIVATION

- In vielen forensischen Szenarien spielen Identitäten eine wichtige Rolle
- So gilt es z.B. für verschiedenste Dokumente zu überprüfen, ob diese von einer verdächtigen Person stammen (oder nicht)
- Zu solchen Dokumenten zählen u.a.:
  - Erpresserschreiben
  - Drohungen
  - Abschiedsbriefe
  - Bekennerschreiben
  - Testamente
  - Schadensmeldungen
  - …

MOTIVATION

- Die Bearbeitung solcher Fälle (Überprüfungen) erfolgt i.d.R. manuell

- Daraus ergeben sich viele Problemstellungen, z.B.:
  - Skalierbarkeit (viel Datenaufkommen, Mehrsprachigkeit, etc.)
  - Kosten für externe Gutachter (einfache Analysen ≈ dreistellige Summen)
  - Zeitlicher Aufwand (Experten benötigen oft Wochen für einen Fall)

- Automatisierte Lösungen existieren, können Experten jedoch nicht ersetzen, vor allem nicht vor Gericht…

- …sollen sie aber auch nicht → stattdessen: Aufwand verringen und legitimen Anfangsverdacht bestätigen

EINFÜHRUNG: AUTORSCHAFTS-VERIFIKATION

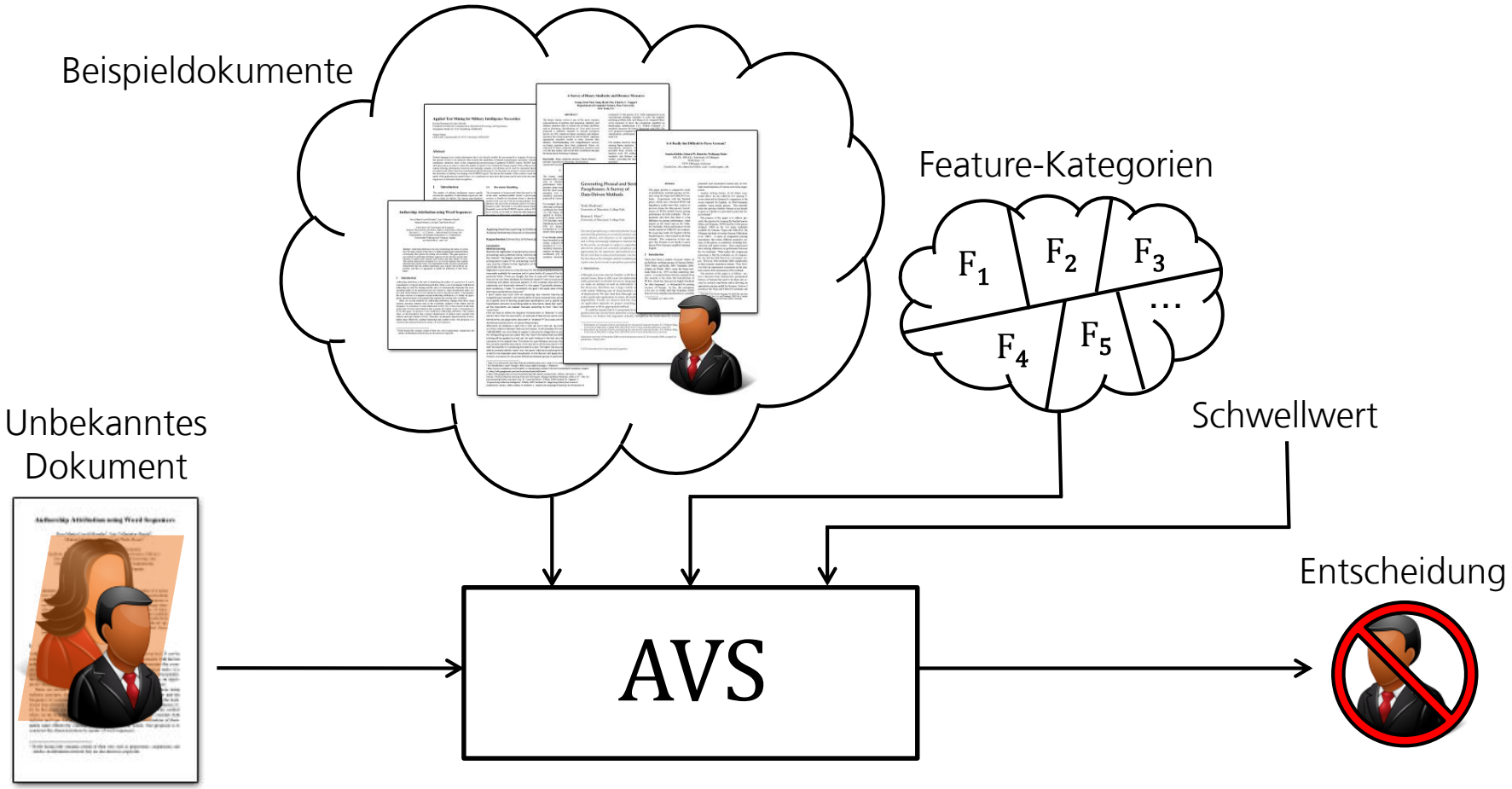
- **Lösung:** Autorschafts-Verifikation

  Verfahren, welches anhand **vorhandener** Beispieltexte einer Person A prüft, ob ein **gegebenes** Dokument ebenfalls von A stammt

- Ein Autorschafts-Verifikationssystem (AVS) ist die Realisierung eines solchen Verfahrens und wurde jüngst von uns entwickelt

- Wie genau sieht ein AVS aus und was setzt dieses voraus?

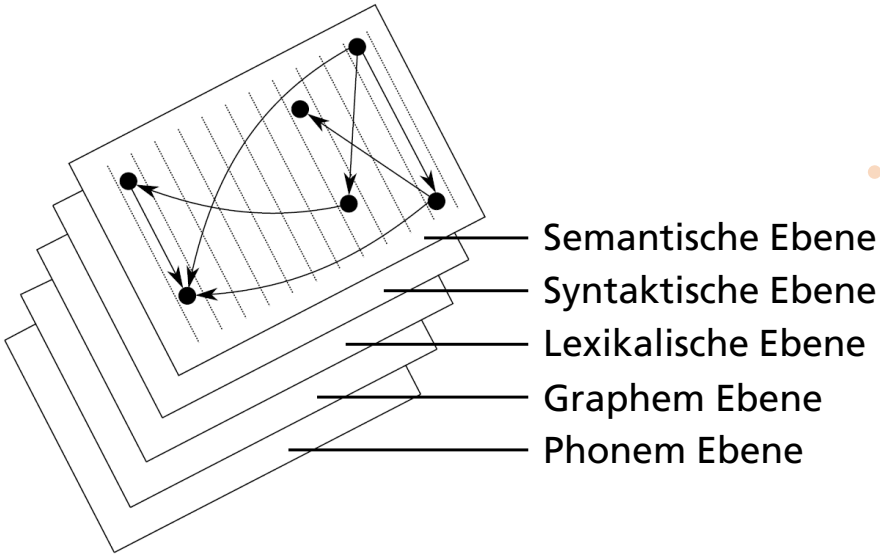
EINFÜHRUNG: AUTORSCHAFTS-VERIFIKATION



FEATURE-KATEGORIEN

- Features (deut. "*stilistische Merkmale*") sind Grundlage eines jedes AVS

- Werden aus den unterschiedlichsten sprachlichen Ebenen eines Textdokuments gewonnen, z.B.



und noch viele weitere…

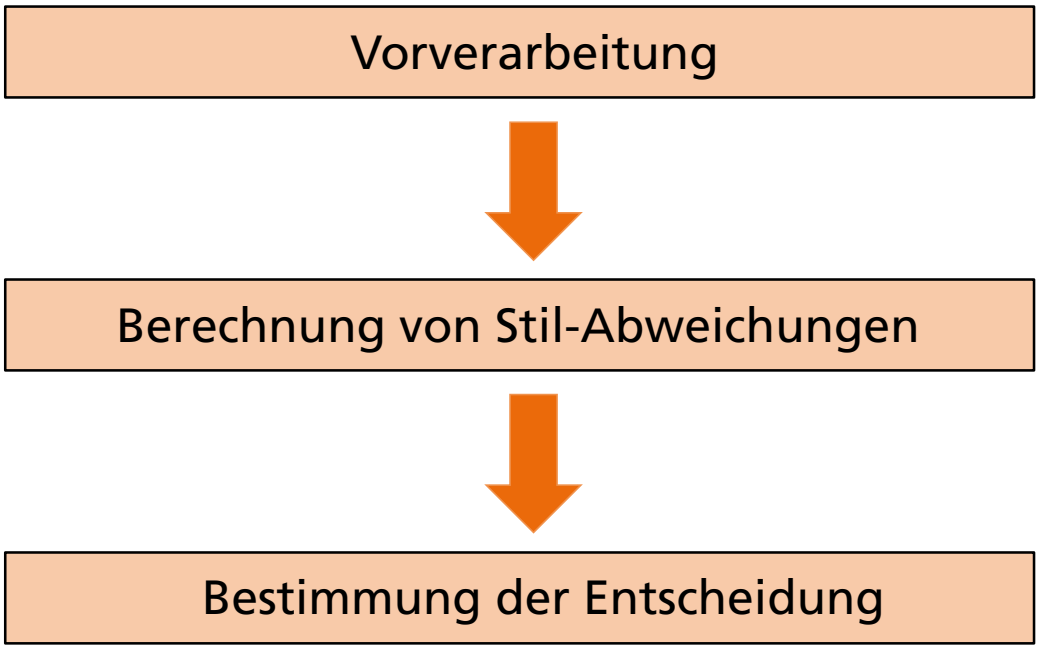
FEATURE-KATEGORIEN

- Eine Feature-Kategorie ist eine Familie von sprachlichen Merkmalen

F_i	Feature-Kategorie	Beispiele
F_1	Interpunktionszeichen	(,), [,], !, ?, ;, :, …
F_2	Buchstaben	A-Z, Ä, Ö, Ü, a-z, ä, ö, ü, ß
F_3	Buchstaben n-Gramme	Dies ist ein Text → (**Di**, **ie**, **es**, **_i**, **is**, …)
F_4	Wort-Präfixe	**Di**es **is**t **ei**n **Te**xt
F_5	Wort-Suffixe	Di**es** i**st** e**in** Te**xt**
F_6	Funktionswörter	und, oder, ist, auch, deswegen, darum, daher, …
F_7	Funktionswort n-Gramme	Dies ist ein Text → (**Dies ist**), (**ist ein**)
F_8	k-beginnende Funktionswörter	Dies ist ein Text → (**Dies**)
F_9	Wort n-Gramme	Dies ist ein Text → (**Dies ist ein**), (**ist ein Text**)
F_{10}	Wort n-Gramme Längen	Dies ist ein Text → (4_3), (3_3), (3_4)
F_{11}	Wort n-Gramme k-Präfixe	Dies ist ein Text → (**Di_is_ei**), (**is_ei_Te**)
F_{12}	Wort n-Gramme k-Suffixe	Dies ist ein Text → (**es_st_in**), (**st_in_xt**)

UNSER VERFAHREN

- Die Prozedur von unserem AVS kann in drei Schritten unterteilt werden



UNSER VERFAHREN / VORVERARBEITUNG

- Beim Vorverarbeiten der Texte wenden wir zum einen eine **Normalisierung** und zum anderen eine **Rauschunterdrückung** an

Wichtig um Texte einheitlich zu behandeln!

→ Substitution von diakritischen Zeichen (ñ → n) oder mehrfache Leerzeichen, etc.

Wichtig um die Qualität der Features zu erhöhen!

→ Entfernung von Zitaten, Markup-Tags, Tabellen, Formeln, Nicht-Wörter, etc.

UNSER VERFAHREN / BERECHNUNG VON STIL-ÄHNLICHKEITEN

- Unser System basiert auf einem bekannten Machine Learning Verfahren

- Dieses setzt voraus, dass sämtliche Texte zunächst in Feature-Vektoren überführt werden:

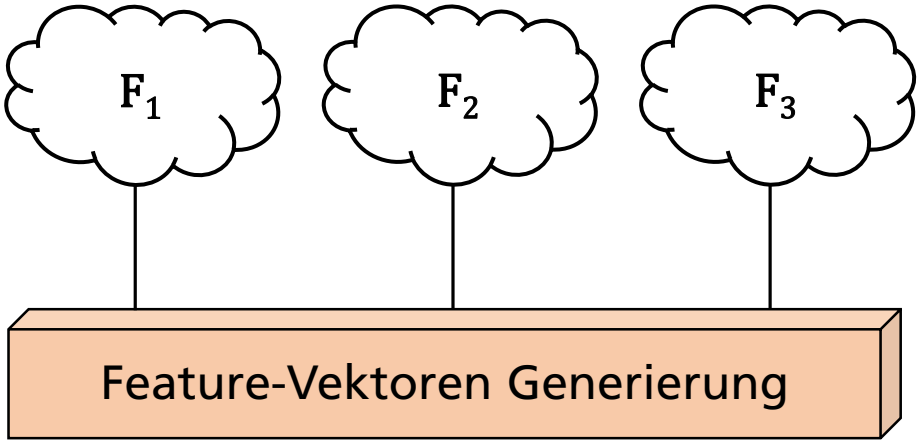
$$Y = (f_1, f_2, f_3, \ldots, f_n)$$

$$X_1 = (f_1, f_2, f_3, \ldots, f_n)$$
$$X_2 = (f_1, f_2, f_3, \ldots, f_n)$$
$$\vdots$$
$$X_m = (f_1, f_2, f_3, \ldots, f_n)$$

$$f_1 = \frac{\text{Anzahl aller "und"}}{\text{Anzahl aller Funktionswörter}}$$

UNSER VERFAHREN / BERECHNUNG VON STIL-ABWEICHUNGEN

- Für jede gewählte Feature-Kategorie werden Vektoren generiert



- Nach der Generierung gilt es Stil-Abweichungen zu berechnen

- Diese werden paarweise für Y und jedes einzelne X_i berechnet

UNSER VERFAHREN / BERECHNUNG VON STIL-ABWEICHUNGEN

- Eine Stil-Abweichung a_i ist ein Wert im Bereich $[0\,;\,\infty)$ und wird anhand einer Distanzfunktion berechnet, z.B. Minkowski:

$$Minkowski(X, Y, \lambda) = \sum_{j=1}^{n} \left(|x_j - y_j|^{\lambda} \right)^{\frac{1}{\lambda}}$$

- **Interpretation:** Je mehr sich a_i an 0 nähert desto ähnlicher ist der Schreibstil zwischen Y und X_i

- Nachdem alle Stil-Abweichungen berechnet wurden, werden diese zusammen mit ihrem zugehörigen Feature-Vektor…

UNSER VERFAHREN / BERECHNUNG VON STIL-ABWEICHUNGEN

- …in eine Liste gespeichert, die absteigend nach den a_i sortiert ist

$$Outer\_Distances = \left((a_1, \mathbf{X_1}), (a_2, \mathbf{X_2}), \ldots, (a_m, \mathbf{X_m}) \right)$$

- Nun wird das erste Tupel entnommen und erneut Stil-Abweichungen berechnet, dieses Mal jedoch zwischen $\mathbf{X_1}$ and $\mathbf{X_2}, \mathbf{X_3}, \ldots, \mathbf{X_m}$

- Auch hier werden die Werte absteigend in einer Liste gespeichert (ohne zugehörige Feature-Vektoren)

$$Inner\_Distances = (a_2, a_3, \ldots, a_m)$$

UNSER VERFAHREN / BESTIMMUNG DER ENTSCHEIDUNG

- Um eine Entscheidung bzgl. einer Feature-Kategorie zu fällen muss der Durchschnitt der k nächsten Nachbarn von a_1 berechnet werden:

$$avg\_kNN = \frac{a_2, a_3, \dots, a_k}{k}$$

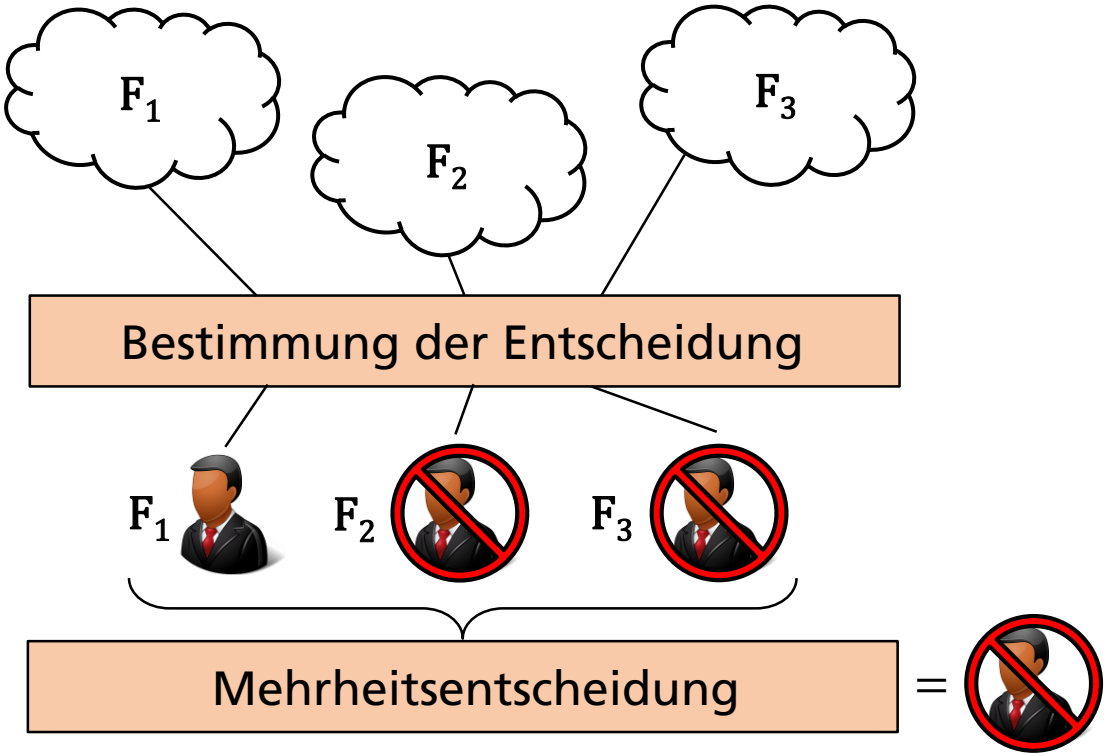
- Anhand von a_1 und diesem Wert wird ein Akzeptanzkriterium bzgl. der vermeintlichen Autorschaft definiert:

$$\frac{a_1}{avg\_kNN} \leq \text{Schwellwert}$$

*In den meisten Fällen ist dieser **1***

UNSER VERFAHREN / BESTIMMUNG DER ENTSCHEIDUNG

- Aus den einzelnen Entscheidungen wird im letzten Schritt schließlich eine Mehrheitsentscheidung durchgeführt



EVALUIERUNG

- Unser System wurde sowohl von uns als auch von unabhängigen Experten eines international ausgerichteten Wettbewerbs evaluiert

PAN - "*Uncovering Plagiarism, Authorship, and Social Software Misuse*"

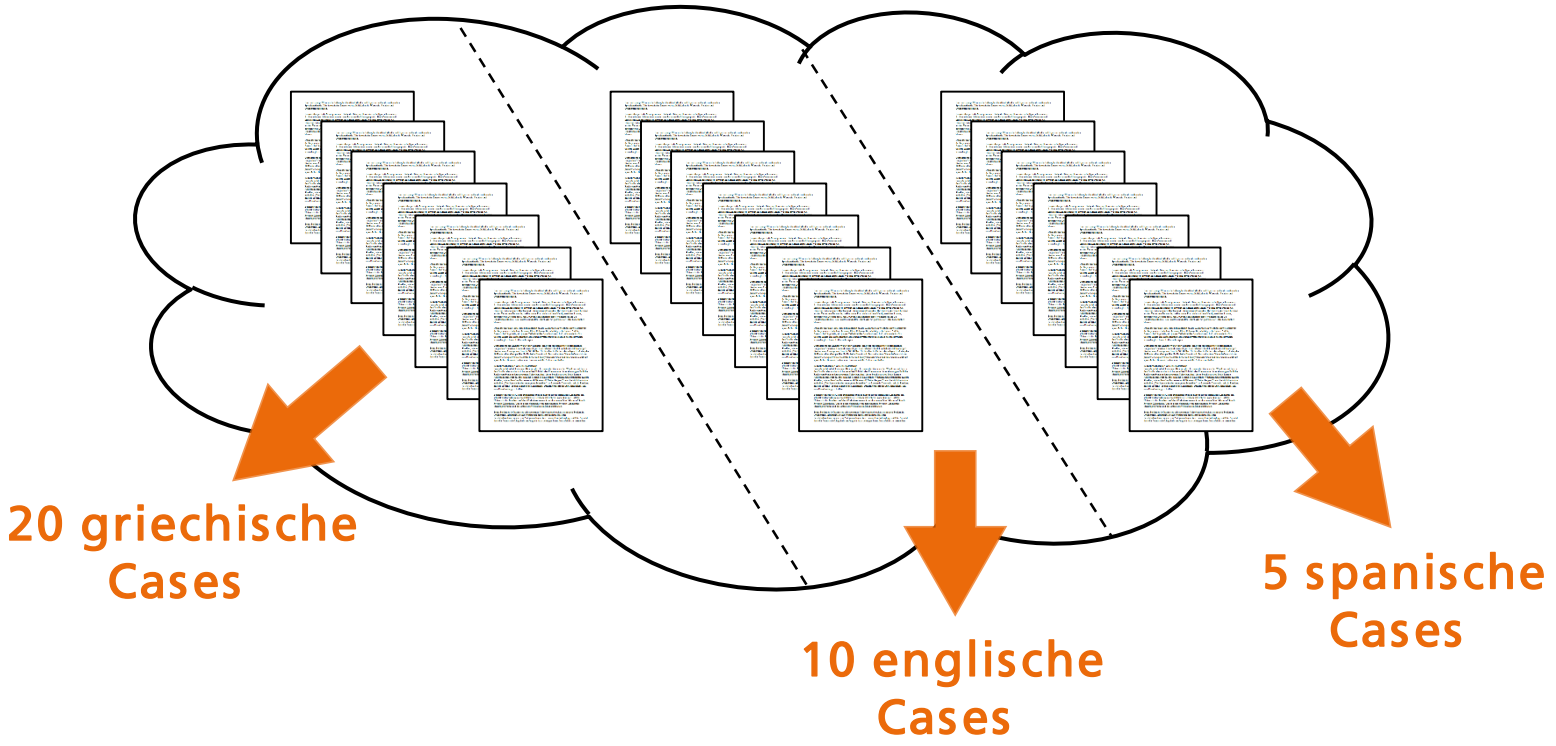
http://PAN.Webis.de

Drei Disziplinen standen 2013 zur Auswahl:

- Plagiarism Detection
- **Author Identification**
- Author Profiling

EVALUIERUNG

- Grundlage der Evaluierung war der öffentlich zugängliche PAN Trainings-korpus mit insg. **189 Dokumente**, aufgeteilt nach „Problem-Cases"



EVALUIERUNG

- Pro Case sind zwischen 4 und 9 Dokumente enthalten
- Davon repräsentiert stets ein Dokument die zu verifizierende Autorschaft und der Rest die Trainingsdaten (Beispieldokumente)
- Jeder Korpus ist bzgl. der Autorschaften balanciert, d.h. ca. 50% der Cases enthalten eine korrekte Autorschaft und der Rest eine falsche
- **Besondere Herausforderung:** Texte wurden so ausgewählt, dass sie absichtlich keine signifikanten Stil-Ähnlichkeit untereinander aufweisen

→ Abschwächung für typische Ansätze in dieser Disziplin ←

EVALUIERUNG

- Ergebnisse bzgl. des PAN Trainingskorpus

$\mathbb{F}$	$\varnothing c_{SP}$	$\varnothing c_{EN}$	$\varnothing c_{GR}$	$\varnothing$	(weighted) $\varnothing$
$\{F_1, F_3, F_9\}$	**80 %**	**90 %**	**70 %**	**80 %**	**77.14 %**
$\{F_1, F_3, F_7, F_8, F_{12}\}$	**80 %**	80 %	65 %	75 %	71.42 %
$\{F_1, F_2, F_3\}$	**80 %**	80 %	55 %	71.67 %	65.71 %
$\{F_1, F_4, F_9\}$	**80 %**	80 %	60 %	73.33 %	68.57 %
$\{F_1, F_3, F_9, F_{11}, F_{12}\}$	**80 %**	80 %	55 %	71.67 %	65.71 %
$\{F_7, F_9, F_{11}\}$	60 %	60 %	50 %	56.67 %	54.28 %
$\{F_3, F_6, F_7, F_{11}, F_{12}\}$	60 %	50 %	55 %	55 %	54.28 %
$\{F_2, F_5, F_6\}$	**80 %**	40 %	40 %	53.33 %	45.71 %
$\{F_3, F_7, F_9\}$	20 %	70 %	50 %	46.67 %	51.43 %
$\{F_4, F_6, F_7\}$	40 %	40 %	60 %	46.67 %	51.43 %

- Die erste Kombination stellt die beste aus insg. **4096** Kombinationen dar

EVALUIERUNG

- Ergebnisse bzgl. des PAN Trainingskorpus + eigenen Korpus, welches aus ings. 40 deutschen Problem-Cases besteht

$\mathbb{F}$	$\varnothing c_{SP}$	$\varnothing c_{EN}$	$\varnothing c_{GR}$	$\varnothing c_{DE}$	$\varnothing$	(weighted) $\varnothing$
$\{F_1, F_3, F_9\}$	**80 %**	**90 %**	**70 %**	67.5 %	**76.86 %**	72 %
$\{F_1, F_3, F_7, F_8, F_{12}\}$	**80 %**	80 %	65 %	**77.5 %**	75.63 %	**74.67 %**
$\{F_1, F_2, F_3\}$	**80 %**	80 %	55 %	75 %	72.5 %	70.67 %
$\{F_1, F_4, F_9\}$	**80 %**	80 %	60 %	62.5 %	70.63 %	65.33 %
$\{F_1, F_3, F_9, F_{11}, F_{12}\}$	**80 %**	80 %	55 %	62.5 %	69.38 %	64 %
$\{F_7, F_9, F_{11}\}$	60 %	60 %	50 %	60 %	57.5 %	57.33 %
$\{F_3, F_6, F_7, F_{11}, F_{12}\}$	60 %	50 %	55 %	62.5 %	56.88 %	58.67 %
$\{F_2, F_5, F_6\}$	**80 %**	40 %	40 %	65 %	56.26 %	56 %
$\{F_3, F_7, F_9\}$	20 %	70 %	50 %	67.5 %	51.86 %	60 %
$\{F_4, F_6, F_7\}$	40 %	40 %	60 %	60 %	50 %	55 %

EVALUIERUNG

PAN 2013

Author Identification

Performances on all test data

Submission	F_1	Precision	Recall	Runtime
seidman13	0.753	0.753	0.753	65476823
halvani13	0.718	0.718	0.718	8362
layton13	0.671	0.671	0.671	9483
petmanson13	0.671	0.671	0.671	36214445
jankowska13	0.659	0.659	0.659	240335
ayala13	0.659	0.659	0.659	5577420
bobicev13	0.655	0.663	0.647	1713966
feng13	0.647	0.647	0.647	84413233
vladimir13	0.612	0.612	0.612	32608
ghaeini13	0.606	0.671	0.553	125655
vandam13	0.600	0.600	0.600	9461
moreau13	0.600	0.600	0.600	7798010
jayapal13	0.576	0.576	0.576	7008
grozea13	0.553	0.553	0.553	406755
gillam13	0.541	0.541	0.541	419495
kern13	0.529	0.529	0.529	624366
baseline	0.500	0.500	0.500	–
petmanson13	0.448	0.700	0.329	20671346
zhenshi13	0.417	0.800	0.282	962598
sorin13	0.331	0.633	0.224	3643942

ZUSAMMENFASSUNG

- Gezeigt wurde ein relativ einfaches aber mächtiges Verfahren zum Zwecke der Verifikation von Autorschaften mit den folgenden Eigenschaften

- **Sprachunabhängig**: Aber nicht **sprachenübergreifend**(!)
  Y als auch sämtliche $X_1, X_2, ..., X_m$ müssen in der selben Sprache vorliegen

- **Geringe Komplexität**: Es werden keine rechenzeitaufwendigen Operationen wie bei anderen Ansätzen benötigt

- **Konfigurierbarkeit**: Im Grunde kann alles ersetzt, modifiziert, kombiniert, ausgetauscht oder erweitert werden

Schwellwert, Distanzfunktion, Feature-Kategorien inkl. Parameter, etc.

HERAUSFORDERUNGEN / AUSBLICK

- **Größte Herausforderung:**
  Sehr viele Parameter-Einstellungsmöglichkeiten!

  Nahezu alles kann parametrisiert werden (Größe der n-Gramme, Anzahl der zu verwendenden Features, verwendete Distanzfunktion, etc.)

- **Mögliche Lösung:**
  Einsatz von genetischen Algorithmen, um zumindest für eine Teilmenge von Feature-Kategorien eine optimale Einstellung zu bestimmen

HERAUSFORDERUNGEN / AUSBLICK

- **Andere Herausforderung:**
  Es ist (noch) nicht geklärt, inwiefern das Thema eines Textes die Ergebnisse beeinflusst, in denen Buchstaben n-Gramme bzw. Wörter involviert sind

- **Vorgehen:**
  Aktuell läuft beim Fraunhofer SIT eine Studie, die dieser Fragestellung anhand von Visualisierungen von Buchstaben n-Grammen in größeren Korpora nachgeht

VIELEN DANK FÜR IHRE AUFMERKSAMKEIT…

Oren.Halvani@SIT.Fraunhofer.de