TeleTrusT
Pioneers in IT security.

TeleTrusT-Informationstag "IT-Forensik"

Berlin, 12.05.2016

Autorschaftsanalyse

Oren Halvani

Fraunhofer-Institut für Sichere Informationstechnologie, Darmstadt

Autorschaftsanalyse

Oren Halvani

Fraunhofer-Institut für Sichere Informationstechnologie, Darmstadt

Fraunhofer SIT

Bundesministerium
für Bildung
und Forschung

CASED

ÜBERBLICK

- Motivation
- Grundlagen der Autorschaftsanalyse
- Unser AV-Verfahren
- Evaluierung
- Beobachtungen, Zusammenfassung & Ausblick

MOTIVATION

- Stetiger Anstieg an Internetbetrug, Identitätsdiebstahl, Hetze, Verleumdung, Informationslecks und weiterer Phänomene

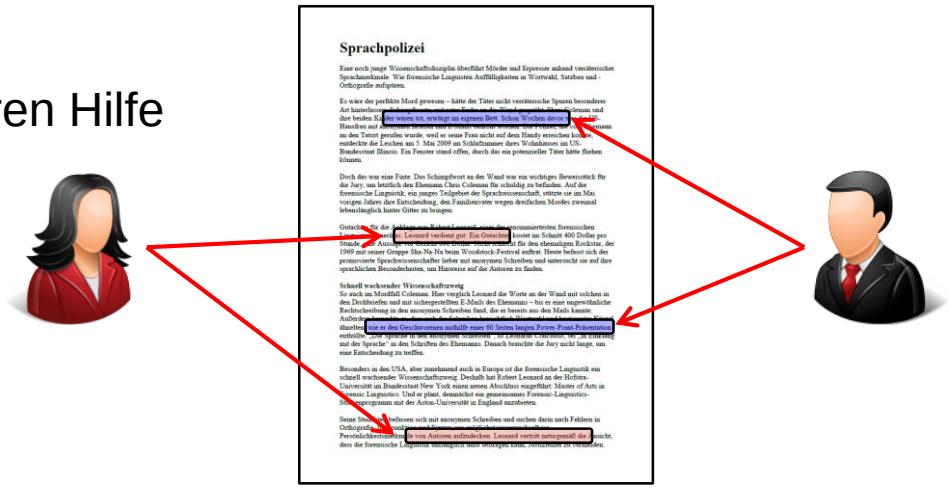
- **Gemeinsamer Nenner:** Hinter jedem Phänomen verbirgt sich mindestens eine Person, die unbemerkt Spuren hinterlässt

- Typische Spuren (IP-Adressen, Log-File-Einträge, etc.) greifen nicht immer bzw. sind nicht verfügbar

MOTIVATION

Wie kann ein Text einem Autor zugeordnet werden?

- Enthält der Text verwertbare Informationen über den Verfasser?
- Texte beinhalten Merkmale mit deren Hilfe sie einem (oder auch mehreren) bekannten Autoren zugeordnet werden können…

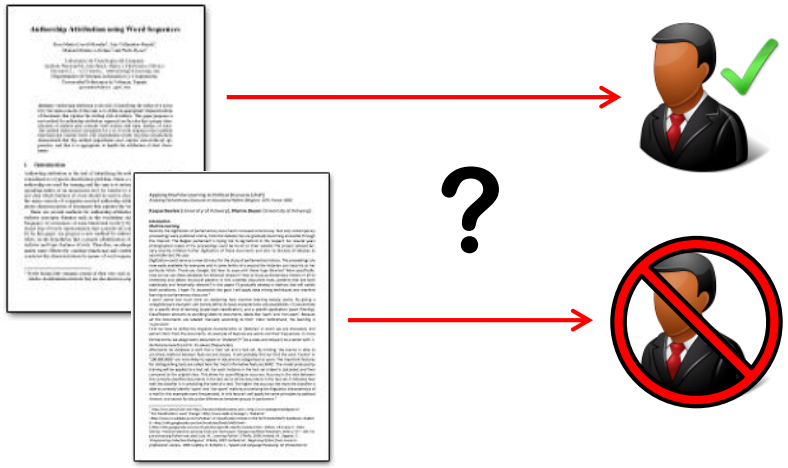


- Die Disziplin die sich mit der Aufdeckung dieser Merkmale auseinandersetzt ist die **Autorschaftsanalyse**

AV → ZIEL

- Fokus dieser Präsentation ist die Autorschaftsverifikation (**AV**), welche die wichtigste Unterdisziplin der Autorschaftsanalyse darstellt

- **Ziel von AV:** Stammen zwei Dokumente von ein und demselben Autor?



AV → ANWENDUNGEN

Erkennung von Versicherungsbetrug

Musterbrief
für Ihre Schadensmeldung

So sieht unser Musterbrief aus

Schadensmeldung
NameAdresse
Straße/Hausnummer
Postfach
Postleitzahl/Ort
Telefon

An die Haftpflichtversicherung
des Unfallverursachers (Name, Anschrift)

Unfall am .....in .....

Sehr geehrte Damen und Herren, durch den genannt
Herr .... als Fahrer des bei Ihnen haftpflichtversiche
verschuldet.
Herr .... ist bei Ihnen unter der Versicherungsnumm
sämtlichen Schadenersatzansprüche aus diesem Un
beziffern. Sobald mir das möglich ist, reiche ich Ihr
Grüßen(Unterschrift)AnlageUnfallberichtUnfallskiz

Note
6,0

Bewertung vom 14.07.2013, Kassenpatient, Alter: 30 bis 50

Was ist das für eine Ärztin?

Ich war vor einer Weile bei Fr. Dr. [redacted] zwecks Untersuchung einiger Muttermale. Ungelogen, die Ärztin wollte mir 3 der Muttermale gleich rausschneiden und ins Labor zur genaueren Untersuchung abschicken. Ich sagte ich werde es mir noch überlegen...ca. 2 Monate später hatte ich einen anderen Hautarzt (in Bad Vilbel) aufgesucht der damals Chefarzt im Uniklinik Frankfurt war. Dieser untersuchte mich gründlich und sagte das die 3 Muttermale völlig unbedenklich sind und das Rausschneiden total überflüssig gewesen wäre. Zum Glück habe ich an meine Zweifel festgehalten !"

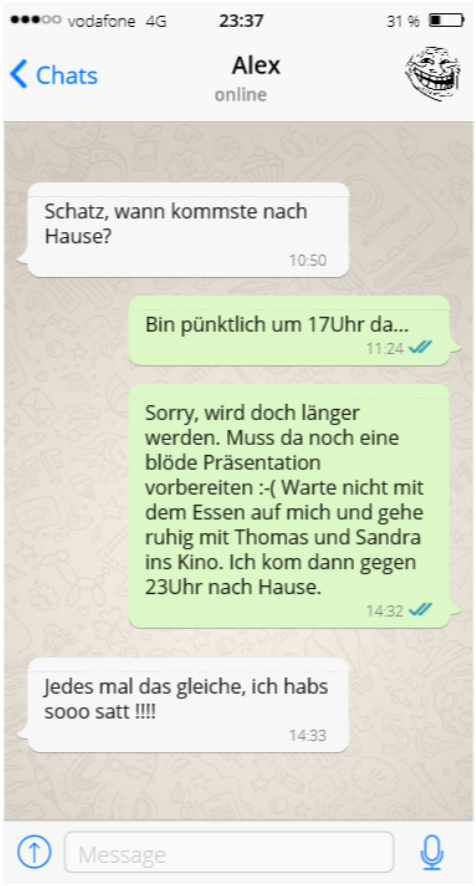
Notenbewertung dieses Patienten

Behandlung	6,0
Aufklärung	6,0
Vertrauensverhältnis	6,0
Genommene Zeit	6,0
Freundlichkeit	6,0

Gesamtnote 6,0

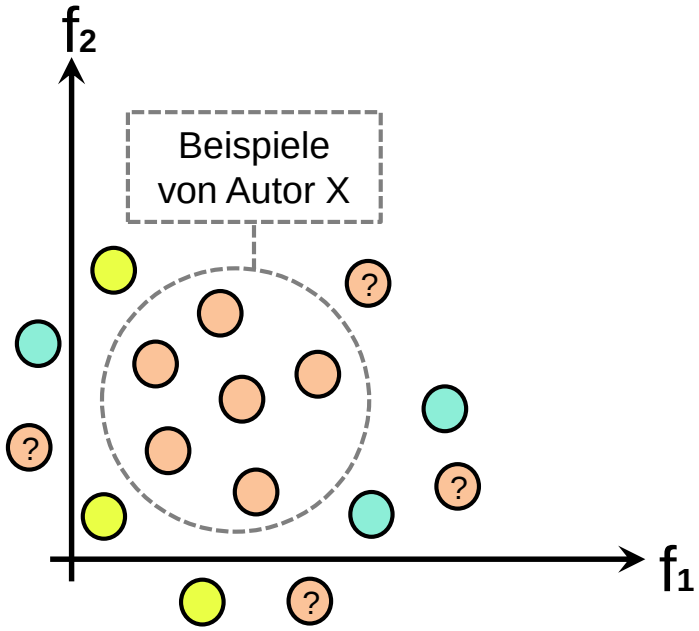
Demaskierung anonymer Bewertungen

Erkennung kompromittierter Accounts



AV → HERAUSFORDERUNGEN

Angenommen ein Fall mit 6 Dokumenten eines Autors X liegt uns vor...



- **Problem 1:** Wie ordnen wir weitere ungesehene Dokumente von X zu?
- **Problem 2:** Es existieren mehrere Milliarden anderer Personen die vorgeben können sie wären ebenfalls X
- **Problem 1 + 2:** Wie können ungesehene Dokumente von X akzeptiert und gleichzeitig andere Autorschaften zurückgewiesen werden?

AV → FEATURES

- Schreibstil ist individuell und damit nicht formalisierbar
- AV erfordert daher eine Approximation des Schreibstils…
- Eine (vielleicht die einzige) Möglichkeit

  → Verwendung stilistischer Merkmale (**Features**)

Features ermöglichen eine Modellierung des Schreibstils

AV → FEATURES

F_i	Name	Feature-Beschreibung	Parameter
F_1	Interpunktionszeichen n-Gramme	Eine Folge von n überlappenden Interpunktionszeichen (Kommas, Bindestriche, etc.)	$n \in \{1, 2, \ldots, 10\}$
F_2	Zeichen n-Gramme	Eine Folge von n überlappenden Zeichen	$n \in \{1, 2, \ldots, 10\}$
F_3	n% häufigsten Token	Die n% häufigsten Tokens (bei $n \approx 30$ sind die Tokens zumeist Funktionswörter).	$n \in \{5, 10, \ldots, 50\}$
F_4	Token k-Präfixe	Die ersten k Zeichen eines Tokens	$k \in \{1, 2, 3, 4\}$
F_5	Token k-Suffixe	Die letzten k Zeichen eines Tokens	$k \in \{1, 2, 3, 4\}$
F_6	Token k-Präfix n-Gramme	Die ersten k Zeichen von jedem Token, innerhalb eines Token n-Gramms	$n \in \{2, 3, 4\}$, $k \in \{1, 2, 3, 4\}$
F_7	Token k-Suffix n-Gramme	Die letzten k Zeichen von jedem Token, innerhalb eines Token n-Gramms	$n \in \{2, 3, 4\}$, $k \in \{1, 2, 3, 4\}$
F_8	n-Präfixe–k-Suffix	Die ersten n und die letzten k Zeichen eines Tokens	$n, k \in \{1, 2, 3, 4\}$
F_9	n-Suffixe–k-Präfixe	Die letzten n Zeichen eines Tokens und die ersten k Zeichen eines Folge-Tokens	$n, k \in \{1, 2, 3, 4\}$

Beispiel für Zeichen 3-Gramme: Halvani → (Hal, alv, lva, van, ani)

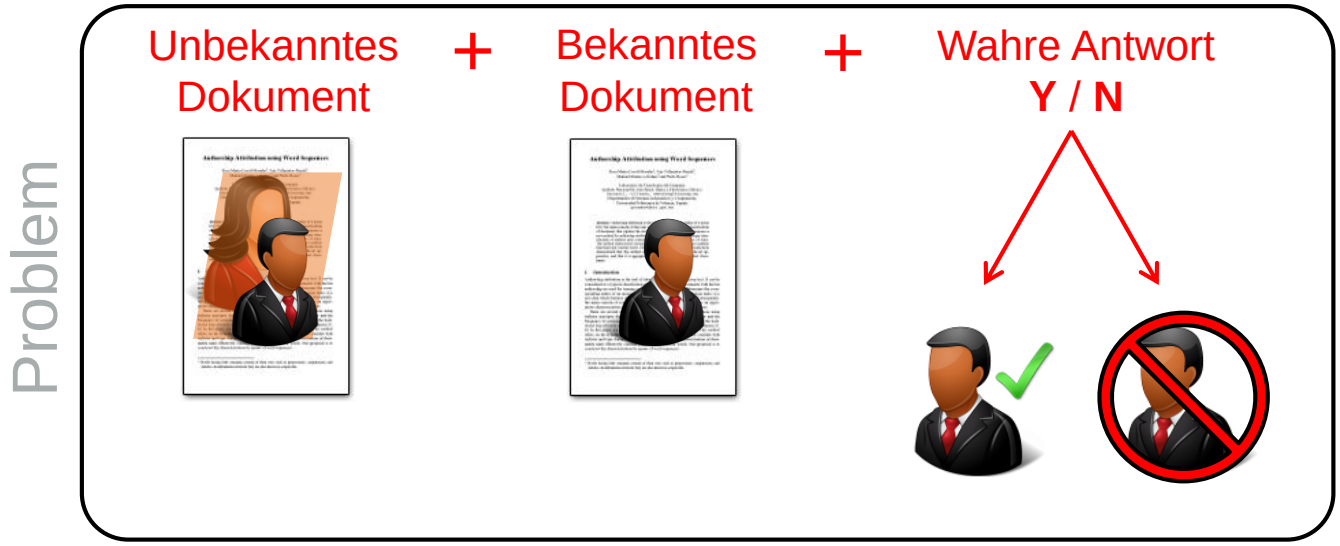
AV → KORPORA

- Um optimale Features und die Güte des AV-Verfahrens zu bestimmen werden Korpora (Dokumentkollektionen) benötigt

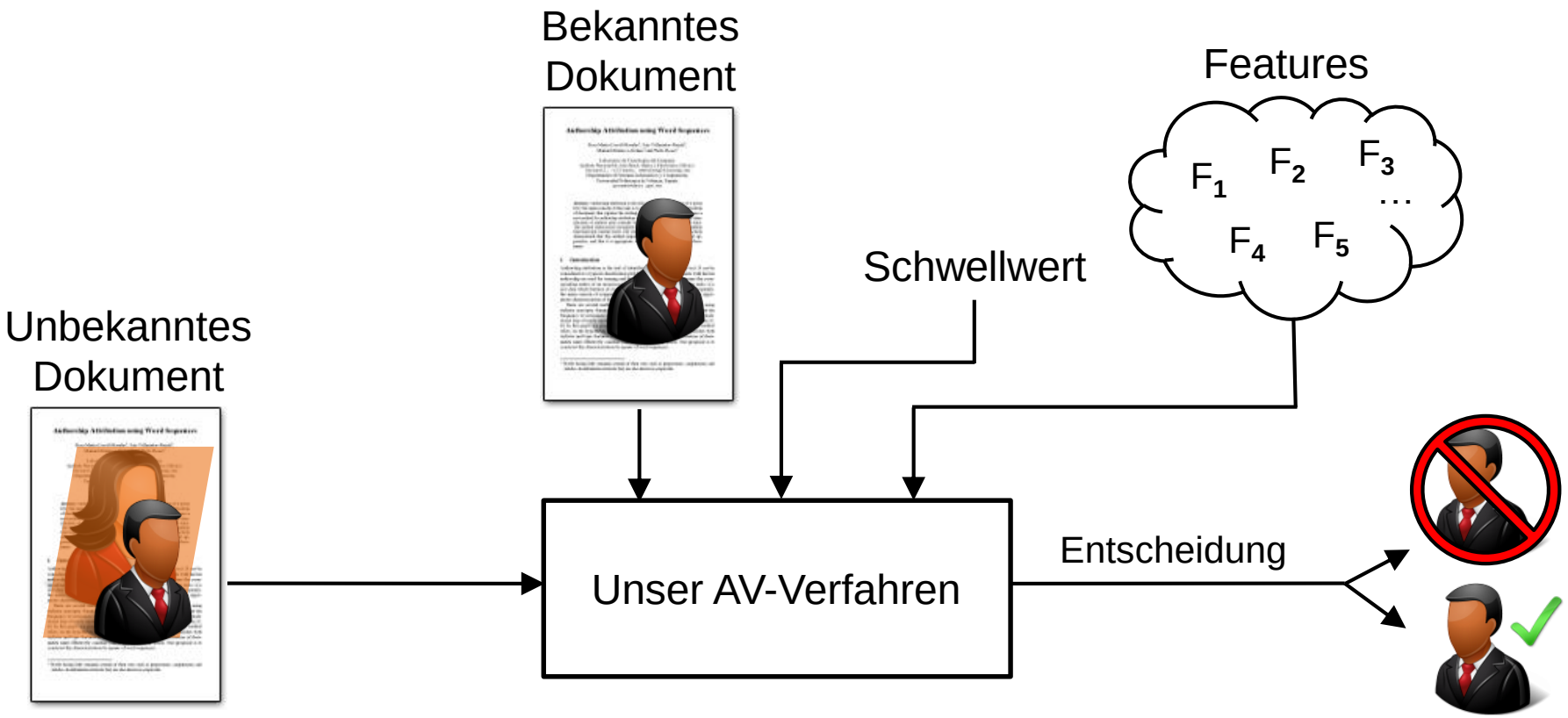
- Für unser AV-Verfahren verwendeten wir zahlreiche Texte aus unterschiedlichen Quellen wie z.B. Nachrichten-Portale, Soziale Netzwerke, Foren, E-Mail-Archive, Abschlussarbeiten, Zeitschriftenartikel, etc.

AV → KORPORA

- Ein Korpus setzt sich aus **n** "Problems" zusammen:



UNSER AV-VERFAHREN



UNSER AV-VERFAHREN

- Generiere für jedes Problem **p** entsprechende Feature-Vektoren:

Bekanntes Dokument → $X = (x_1, x_2, x_3, \ldots, x_n)$

Anzahl aller Der‘s im Text / Anzahl aller Wörter im Text

Unbekanntes Dokument → $Y = (y_1, y_2, y_3, \ldots, y_n)$

UNSER AV-VERFAHREN

- Berechne Ähnlichkeitswerte:

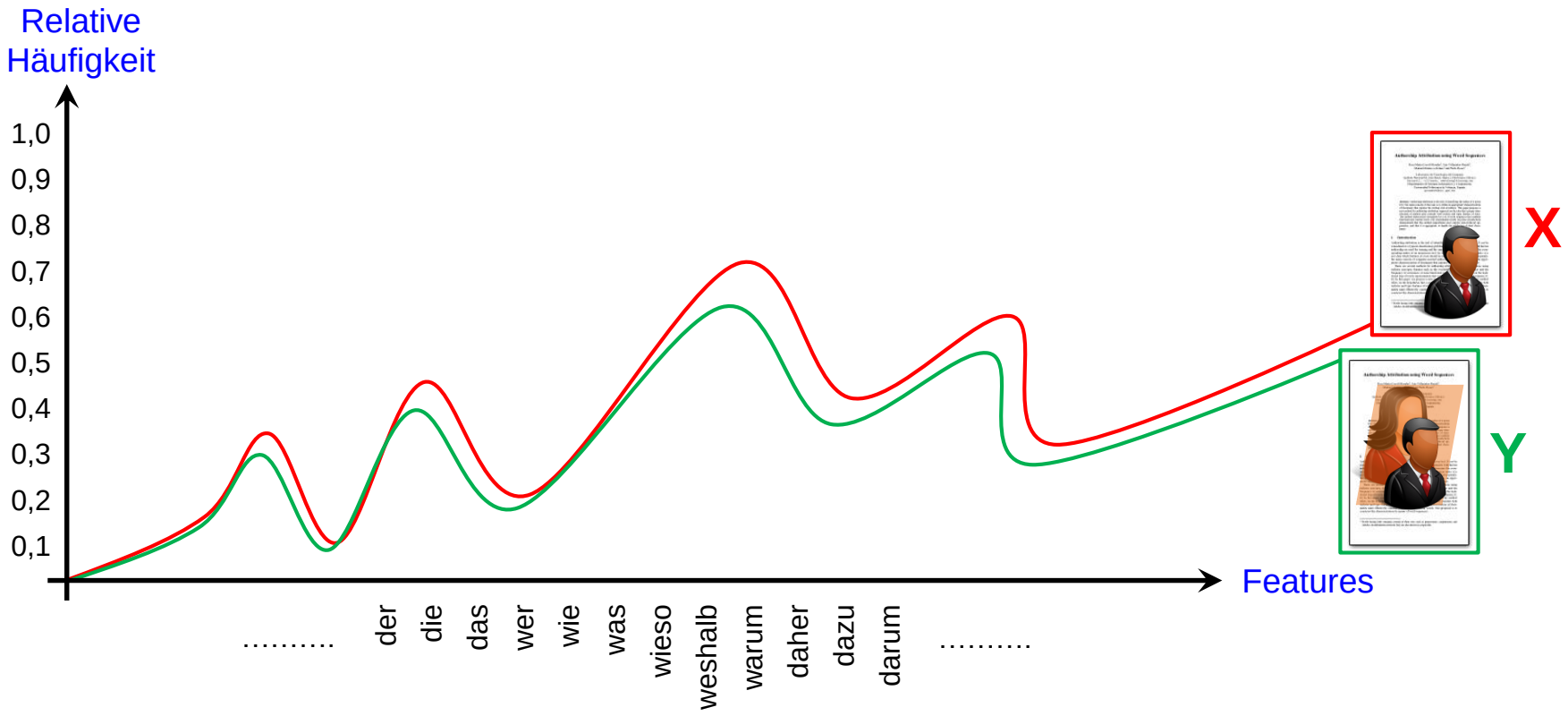
$$\mathrm{sim}(X,Y) = \frac{1}{1+\sum_{i=1}^{n}|x_i - y_i|}$$

- Klassifiziere schließlich jedes Problem **p** hinsichtlich dessen Autorschaft:

$$\mathrm{classify}(\rho) = \begin{cases} \texttt{Y} & \text{falls } \mathrm{sim}(X,Y) > \text{Schwellwert} \\ \texttt{N} & \text{sonst} \end{cases}$$

UNSER AV-VERFAHREN

- **Ziel:** Suchen nach ähnlicher Wahrscheinlichkeitsdichte !



EVALUIERUNG

- Unser AV-Verfahren wurde zunächst auf 28 Test-Korpora evaluiert

  Statistik: 4.525 Problems, verteilt über fünf Sprachen, 16 Genres und > 1.000 gemischte Themen

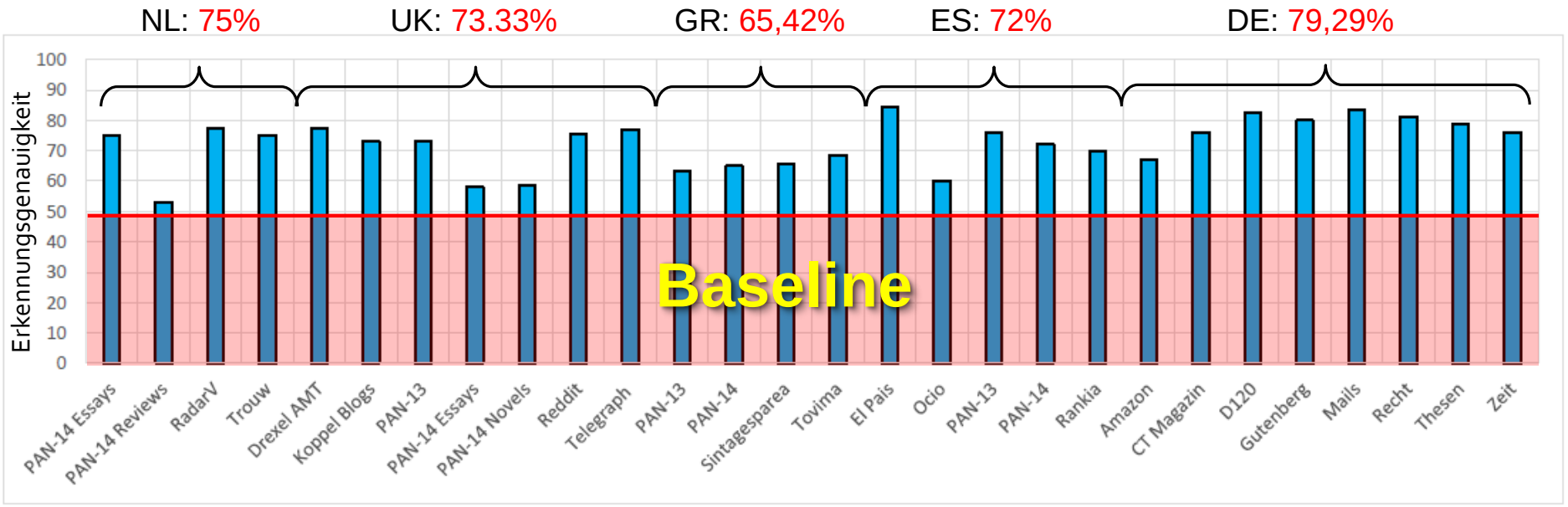
- Anschließend wurde unser AV-Verfahren gegen Verfahren von zwei anderen führenden Forschern (Stamatatos und Moreau) verglichen

- Weiterhin wurde unser AV-Verfahren, im Rahmen des international ausgerichteten PAN-Wettbewerbs gegen 17 Teilnehmer verglichen

  http://pan.webis.de

EVALUIERUNG

- Ergebnisse der Evaluierung bezüglich der 28 Test-Korpora:



- Erkennungsgenauigkeit, in Median:
  75% (unser AV-Verfahren), 69.3% (Stamatatos), 70% (Moreau),

EVALUIERUNG (PAN 2015)

- Resultat bzgl. des PAN-2015 Wettbewerbs (Evaluierung auf 1.265 Problems)
- Performanz ist hier auf eine Maßzahl zwischen 0 und 1 festgelegt
- **Beobachtung:** Unser AV-Verfahren verhält sich sprachneutral gegenüber den meisten anderen Ansätzen

		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

Quelle: "PAN15-AI-Overview Paper"

BEOBACHTUNGEN

- AV benötigt ausreichende Textlänge (min. 5 KByte ≈ 1 DIN-A4 Seite) um brauchbare Resultate zu erzielen
- AV funktioniert auch dann, wenn Dokumente aus unterschiedlichen Genre und Themengebiete stammen
- 5% der häufigsten Features innerhalb der Texte machen den primären Teil des Schreibstils aus
- Zeichen n-Gramme sind als Features ungeschlagen, allerdings ist es noch unklar was dazu führt und ob sich diese tatsächlich Genre- bzw. Themen-neutral verhalten…

ZUSAMMENFASSUNG

Einige Vorteile unserer Methode:

- **Universell:** Anwendbar auf Englisch, Deutsch, Niederländisch, Spanisch, Griechisch (aber auch Polnisch, Französisch und Schwedisch)
- **Unabhängig & transparent:** Keine linguistischen Ressourcen oder sonstige 3rd-Party Bibliotheken → Einfache Re-Implementierung !
- **Schnell:** Der eingesetzte Algorithmus ist sehr schnell, ein einziges Problem lässt sich nahezu in Echtzeit verifizieren

AUSBLICK

- Um gerichtlich verwertbar zu sein, müssen die Ergebnisse des AV-Verfahrens durch Menschen nachvollziehbar sein

- Aktuell nutzen wir nur reine Häufigkeitsverteilungen um Schreibstile zu modellieren. Ein Schreibstil ist jedoch deutlich komplexer aufgebaut…

- Texte sind i.d.R. nicht statisch → Wir forschen daran unser AV-Verfahren robust gegenüber Textmodifikationen zu gestalten…

Vielen Dank für ihre Aufmerksamkeit !

AUSFÜHRLICHE STUDIE

Digital Investigation 16 (2016) S33–S43

Contents lists available at ScienceDirect

Digital Investigation

journal homepage: www.elsevier.com/locate/diin

DFRWS 2016 Europe — Proceedings of the Third Annual DFRWS Europe

Authorship verification for different languages, genres and topics

Oren Halvani*, Christian Winter, Anika Pflug

Fraunhofer Institute for Secure Information Technology SIT, Rheinstr. 75, 64295 Darmstadt, Germany [1]

http://www.sciencedirect.com/science/article/pii/S1742287616000074