

# Autorschaftsanalyse

Oren Halvani,  
Fraunhofer SIT, Media Security & IT Forensics

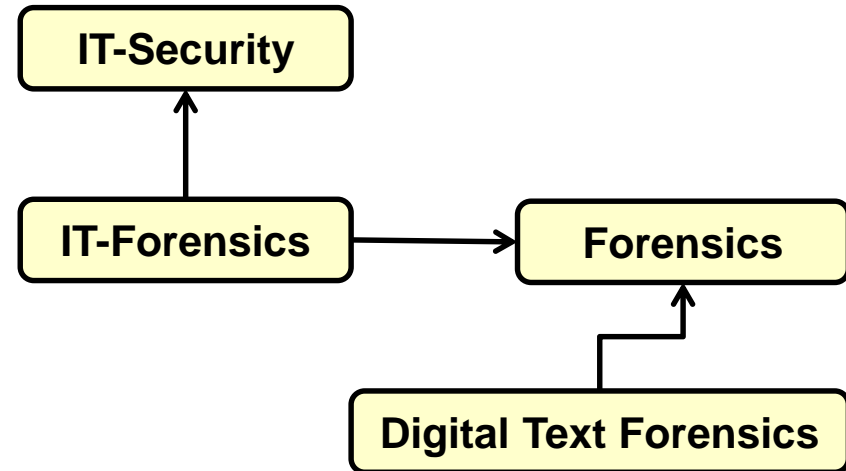
For this reason the amendment should read, or its Credit Support Provider shall fail to have a rating from either S P or Moody's. Please confirm that this is your understanding and sorry for the confusion. David Runnells has come up with a lawyer from A K to help us out. have worked with him and Shonnie left me a voice mail telling me that he was very good. Allegheny energy - Patricia Clark sent to Frank Davis her comments on June 15th by e-mail. I responded on June 19th and have never heard back from her. Most of her comments we could not accept. None of her comments have been run by Dave. Imperial Oil - Grant Oh was handling this. I spoke with Grant about this around June 6th and have never heard back from him. None of their proposed changes have been run by Dave. Kennecott Coal Sales Company - A draft amendment letter which Dave F. has approved was sent to them on May 15th. My contact there is Jim Sobule Wayne Gresham is also involved in this. Arco - We received by e-mail comments from David Dyck at Arco in March. Tori Kuykendall and Grant Oh are involved from the Enron side. I responded to David Dyck by e-mail on March 10th and have never heard anything further on this. None of their comments or my response has been reviewed by Dave. Chevron - Canada - I received a letter from Bruce Borwick on January 31st which I responded to and have never heard anything further from them. Dave has not reviewed their comments or my response. I sent to Mike Pederson a form of amendment letter which I don't think has been signed. They wanted a minor change to the confidentiality language. Dave has not seen this but in the event that Cargill comes back to life, I doubt that he will have a problem with it. Cinergy - I sent a draft amendment letter which Dave approved to John Dirheimer on May 15th. Duke Energy - I sent a draft amendment letter to John Puett on May 15th which Dave approved. To my knowledge, it has not been signed. We are close to reaching agreement with them and mark has been involved in some of the early discussions. I am sending to Dave a mark-up of the revised agreement for his approval. I'm just copying you on these types of memos in case for some reason people come to you with questions. Status of Online Amendments Do I need to follow up with her or what? Allegheny energy - Patricia Clark sent to Frank Davis her comments on June 15th by e-mail. I responded on June 19th and have never heard back from her. Most of her comments we could not accept. None of her comments have been run by Dave. Imperial Oil - Grant Oh was handling this. I spoke with Grant about this around June 6th and have never heard back from him. None of their proposed changes have been run by Dave. Kennecott Coal Sales Company - A draft amendment letter which Dave F. has approved was sent to them on May 15th. My contact there is Jim Sobule Wayne Gresham is also involved in this. Arco - We received by e-mail comments from David Dyck at Arco in March. Tori Kuykendall and Grant Oh are involved from the Enron side. I responded to David Dyck by e-mail on March 10th and have never heard anything further on this. None of their comments or my response has been reviewed by Dave. I received a letter from Bruce Borwick on January 31st which I responded to and have never heard anything further from them. Dave has not reviewed their comments or my response. Here is my first cut at a proposed form of amendment. I have added a rep and an additional Event of Default to cover the 'Lien' issue in light of the fact that Barrett's Credit Agreement restricts them from entering into any agreement that contains a negative pledge. I winged the amendment to the Confirmations since I have not seen them yet.

## It's all about style...

- Motivation
- Schreibstil
- Digital Text Forensics
- Autorschaftsverifikation
- COAV / OCCAV
- Fragen / Diskussion

# Wer bin ich?

- 2007 – 2012: Informatikstudium mit Schwerpunkt ML, NLP, IT Security
- Seit 2012 WiMi am Fraunhofer SIT (Abteilung: Media Security & IT Forensics) mit dem Forschungsschwerpunkt **"Digital Text Forensics"**.
- Ein interdisziplinäres Feld das primär Informatik und Linguistik mit einer Prise Psychologie verbindet...



# Motivation

# Motivation

- Mit zunehmender Digitalisierung in unserer heutigen Zeit entstehen neue Wege der Kommunikation auf Basis von Sprach-, Bild- oder Videobotschaften...
- Trotz dieser zählen **Texte** jedoch nach wie vor zu den primären Informationsträgern.
- Texte enthalten gegenüber anderer Dateiformate keinerlei Meta-informationen, die eine nähere Auskunft über ihren **Ursprung** preisgeben...

20180523\_101007.jpg

IrfanView JPG File



Aufnahmedatum:	23.05.2018 10:10
Markierungen:	Markierung hinzufügen
Bewertung:	☆☆☆☆☆
Abmessungen:	3024 x 4032
Größe:	3,24 MB
Titel:	Titel hinzufügen
Autoren:	Autor hinzufügen
Kommentare:	Kommentare hinzufügen
Kamerahersteller:	samsung
Kameramodell:	SM-G950F
Betreff:	Betreff angeben
Blendenzahl:	F/1.7
Belichtungszeit:	1/984 Sek.
ISO-Filmempfindlichkeit:	ISO-40
Lichtwert:	0 Schritt(e)
Brennweite:	4 mm
Maximale Blende:	1.53
Messmodus:	Mittenbetont
Blitzlichtmodus:	Ohne Blitzlicht
35mm Brennweite:	26
Erstelldatum:	27.05.2018 14:11
Änderungsdatum:	23.05.2018 10:10

- In manchen Szenarien ist aber genau die Frage nach dem Ursprung (**Autorschaft**) bestimmter Texte von großer Bedeutung (z.B. Testamente, Hetzbotschaften, Bekennerschreiben, Verleumdungen, Drohungen oder auch **Abschlussarbeiten** 😊)
- Weiterhin existieren Szenarien in denen sich die Frage stellt, ob die Autorschaft von Texten über eine Zeitspanne hinweg konsistent bleibt.
- Dazu zählen etwa kompromittierte Accounts (z.B. E-Mail-Konten, Social-Media-Plattformen, Instant-Messenger, etc.) von denen kontinuierlich Texte abfließen...

Welche **Person** steht hinter dem Text?

- Tatsächlich ist die Frage hinsichtlich der Autorschaft eines Textes in der Praxis häufig anzutreffen.
- Betrachtet man dort die jeweiligen Szenarien in Detail, so lassen sich unterschiedliche Autorschaftsprobleme identifizieren:
  - **Fehlende** Autorschaften
  - **Anonyme** Autorschaften
  - **Nicht-eindeutige** Autorschaften

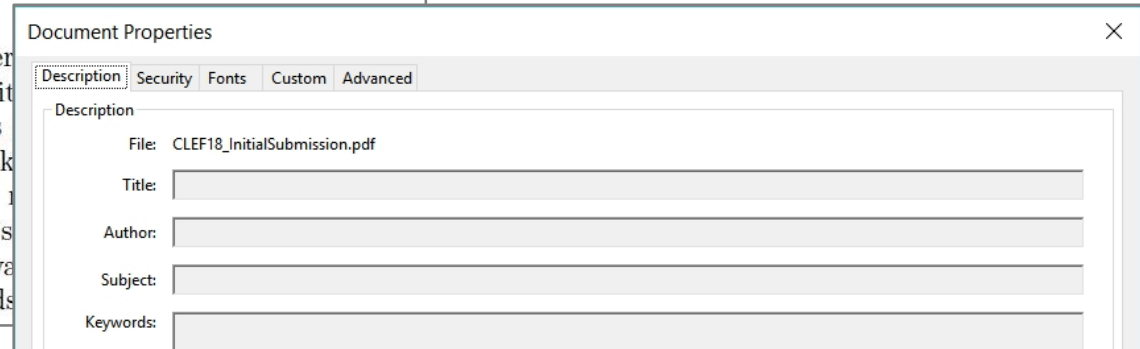
## Rethinking the Evaluation Methodology of Authorship Verification Methods

Authors removed for peer review

No Institute Given

**Fehlende Autorschaften**

**Abstract.** Authorship verification (AV) concerns the task of judging, if two or more documents have been written by the same author. Even though an increase of research activities in this field has been observed, it can also be clearly seen that AV lacks standardized standards. Based on a review of more than 50 peer-reviewed conference papers, journals, bachelor's/master's theses and dissertations, we could not identify consistent evaluation metrics that adequately reflect the reliability of AV methods.



Document Properties

Description Security Fonts Custom Advanced

Description

File: CLEF18\_InitialSubmission.pdf

Title:


Author:

Subject:

Keywords:



## Anonyme Autorschaften



**D120.de/forum**  
Fachschaft Informatik  
FB Informatik  
TU Darmstadt

Schnellzugriff [FAQ](#) [Regeln](#)

Foren-Übersicht < Wahlbereich < Web, Wissens- und Informationsverarbeitung < Einführung in die Künst...

### Backward Search (Inverse Action Application)

Moderator: Einführung in die Künstliche Intelligenz

[Antworten](#) [↩](#) [🔧](#) [⌵](#)  [🔍](#) [⚙️](#)

**Backward Search (Inverse Action Application)**  
von [mProg](#) 11. Jun 2018 19:46

Ich habe eine Verständnisfrage bzgl. Inverse Action Application. Anhand der Folien kann man Backward Search anwenden wenn folgende Bedingung gilt:

how does the hyper-plane is being chosen so that indeed all points that are X(what is X?) far from the hyper-plain are being classified as outliers.

[svm](#) [anomaly-detection](#) [one-class](#)

[share](#) [cite](#) [edit](#)

asked 39 mins ago

[DsCpp](#) 1

★★★★★ **Gerne wieder**

Von [Amazon Customer](#) am 30. Juni 2018

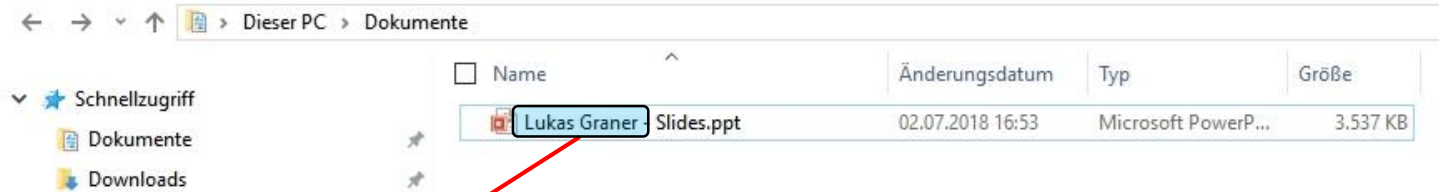
Stil: Deutsche Version | **Verifizierter Kauf**

Hat alles gut geklappt

[Nützlich](#) [Nicht nützlich](#) | [Kommentar](#) | [Missbrauch melden](#)

# Motivation

## Nicht-eindeutige Autorschaften



Max Mustermann - Slides.ppt  
Microsoft PowerPoint 97-2003-Präsentati...



Änderungsdatum: 02.07.2018 16:53  
Autoren: **Martin Steinebach**  
Markierungen: Markierung hinzufügen...  
Größe: 3,45 MB  
Titel: Autorschaftsanalyse  
Kommentare: ii  
Kategorien: Digitale Text Forensik  
Folien: 34  
Inhaltstatus: Geheim ;-)  
Inhaltstyp: application/vnd.ms-...  
Betreff: Betreff angeben  
Erstelldatum: 02.07.2018 16:48



**Autorschaftsanalyse**

Oren Halvani,  
Fraunhofer SIT, Media Security & IT Forensics

TECHNISCHE  
UNIVERSITÄT  
DARMSTADT

02.07.2018 | Fachbereich 20 | Security in Information Technology | Oren Halvani | 1

Fraunhofer  
SIT

It's all about style...

- **Erkenntnis:** Metadaten-basierte Autorschaften sind nicht verlässlich!
- **Abhilfe:** Metadaten ignorieren und stattdessen den Text direkt betrachten...

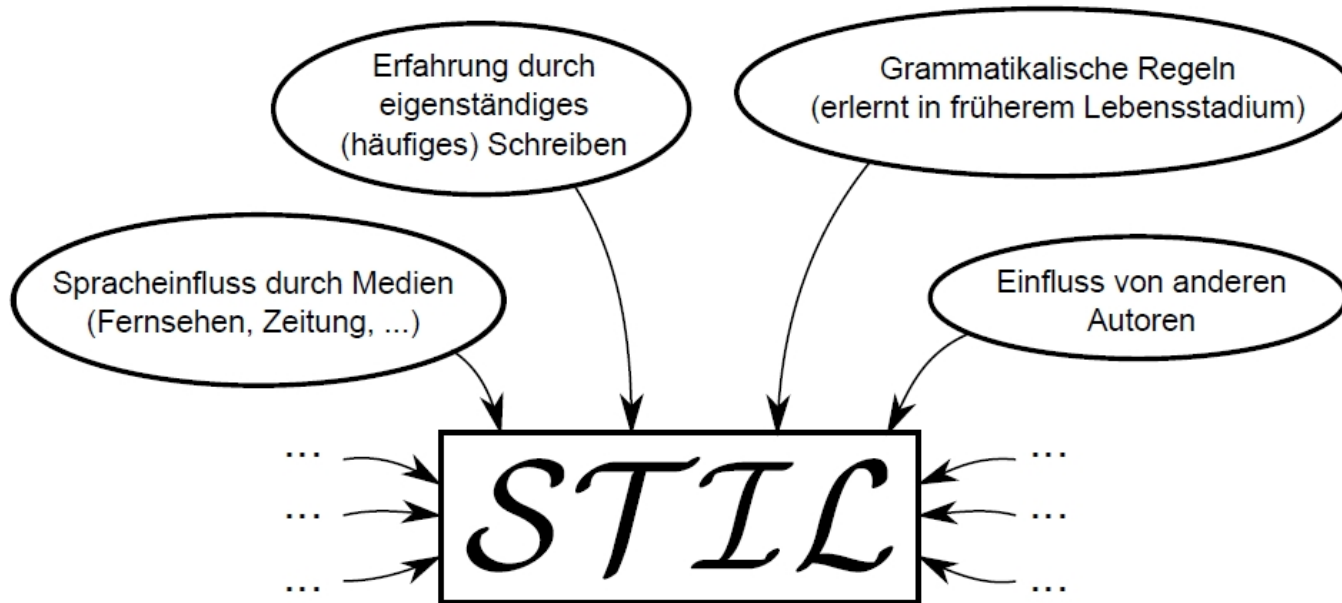


- Nahezu jeder Text beinhaltet ab einem gewissen Umfang einen erkennbaren **Schreibstil**.
- Dieser lässt sich nicht ohne Weiteres anonymisieren oder gar verstecken.

# Schreibstil

- Mit zunehmender Textlänge ist ein **individueller Schreibstil** alleine schon mit "bloßen Auge" beobachtbar.
- Mithilfe statistischer Methoden kann dieser zudem quantifizierbar und damit fassbar gemacht werden.
- Je nach Textsorte ist es leicht bis sehr schwer die individuelle Note eines Autors zu erkennen. Während in **narrativen** Texten (z.B. Romane) der Schreibstil im Vordergrund steht, ist dieser in **normativen** Texten (z.B. Gesetzestexte oder technische Erläuterungen) nahezu nicht vorhanden.

- Der Schreibstil einer Person ist i.d.R. keine statische Eigenschaft...



- Allerdings gibt es bestimmte sprachliche Muster, die selbst über Jahrzehnte hinweg beibehalten werden. Was aber macht genau den Schreibstil einer Person aus?
  - Der **Wortschatz** des Autors?
  - Individuelle **Rechtschreibfehler** die markant sind für einen Autor?
  - Bestimmtes Repertoire an **Funktionswörtern** in einem Text?
  - Bestimmte **Anordnung von Wörtern** innerhalb der Sätze?
  - Die Textkomplexität (**Wortlänge**, **Satzlänge**, durchschnittliche **Vokalanzahl** in Wörtern)?
  - Oder wird Stil nur durch **rhetorische Stilmittel** beschrieben?

- Zählt das als Schreibstil?

$\mathcal{T} =$  „*Op*a *Rainer* hat *seinem Enkel das Buch gestern* *geschenkt.*“

$\mathcal{T}'_1 =$  „*Op*a *Rainer* hat  $\boxed{\sigma_1}$   $\boxed{\sigma_3}$   $\boxed{\sigma_2}$  *geschenkt.*“

$\mathcal{T}'_2 =$  „*Op*a *Rainer* hat  $\boxed{\sigma_2}$   $\boxed{\sigma_1}$   $\boxed{\sigma_3}$  *geschenkt.*“

$\mathcal{T}'_3 =$  „*Op*a *Rainer* hat  $\boxed{\sigma_2}$   $\boxed{\sigma_3}$   $\boxed{\sigma_1}$  *geschenkt.*“

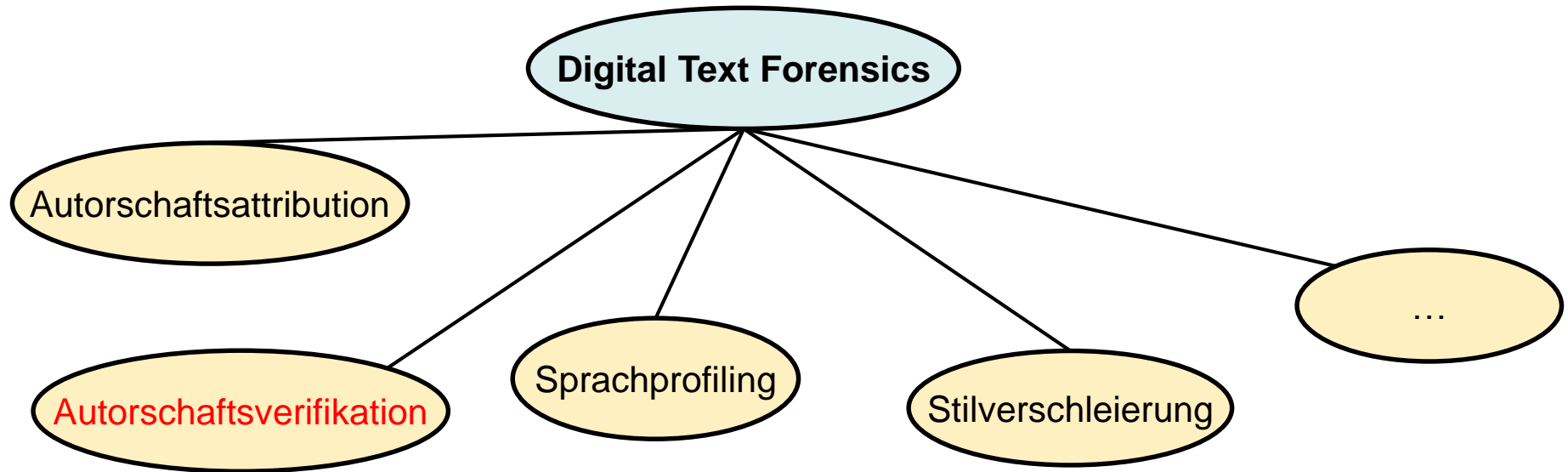
$\mathcal{T}'_4 =$  „*Op*a *Rainer* hat  $\boxed{\sigma_3}$   $\boxed{\sigma_1}$   $\boxed{\sigma_2}$  *geschenkt.*“

$\mathcal{T}'_5 =$  „*Op*a *Rainer* hat  $\boxed{\sigma_3}$   $\boxed{\sigma_2}$   $\boxed{\sigma_1}$  *geschenkt.*“



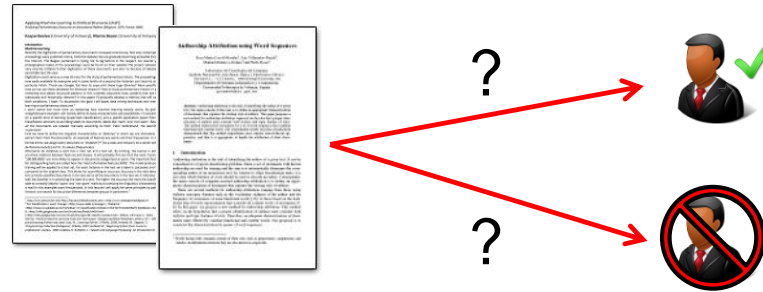
# Digital Text Forensics

- Digital Text Forensics ist ein junges Forschungsfeld das sich in mehrere Unterdisziplinen aufspaltet...



- **Autorschaftsattribute:** Zuordnung von Autoren zu anonymen Texten.
- **Autorschaftsverifikation:** Überprüfung vermeintlicher Autorschaften.
- **Sprachprofiling:** Bestimmung von Eigenschaften von Autoren.
- **Textwiederverwendung:** Erkennung "geborgener" Textpassagen.
- **Autorverschleierung:** Gegenstück aller oben genannten Disziplinen.

- Die Autorschaftsverifikation (**AV**) ist diejenige Disziplin, die sich mit der Konsistenz von Schreibstilen beschäftigt. Ihr Ziel ist die Bestimmung, ob zwei oder mehr Texte von ein und demselben **generierenden Prozess** (i.d.R. ein Autor) stammen.



- Es handelt sich hierbei um ein Ähnlichkeitsproblem, welches die stilistische Ausprägung des Textes statt dessen Inhalt betrachtet.

- Was wir **automatisiert** zu lösen versuchen ist der folgende Verifikationsfall:



# Häufige Vorgehensweise

- Existierende AV-Verfahren benötigen eine spezifische **Repräsentation** der zu untersuchenden Texte (Bag-of-Features / Vector-Space Models, Language Models, Embeddings, etc.)
- Basis dieser Repräsentationen sind Features (stilistische Merkmale), die automatisiert oder manuell aus den Texten extrahiert werden müssen. Letzteres benötigt **Domänenwissen**:
  - Welche Features eignen sich für welche Textsorte und Sprache?
  - Wie müssen parametrisierte Features konfiguriert werden?
  - Wie viele Features sollen betrachtet werden?
  - Wie müssen Features normalisiert werden?



# Autorschaftsverifikation

- Eine kleine Auswahl von Features...

Feature-Kategorie	Kurzbeschreibung / Beispiele
Interpunktionszeichen	(, ), [ , ], !, ?, ;, :, ...
Buchstaben	A-Z, Ä, Ö, Ü, a-z, ä, ö, ü, ß
Buchstaben n-Gramme	Textbeispiel $\xrightarrow{n=2}$ {Te, ex, xt, tb, be, ...}, $\xrightarrow{n=3}$ {Tex, ext, xtb, tbe, ...}, ...
Präfixe	<b>Text</b> beispiel (Vorsilbe)
Infixe	Text <b>be</b> ispiel (innerer Wortbestandteil)
Suffixe	Textbe <b>piel</b> (Nachsilbe)
Funktionswörter	Artikel (der, das, einer, eines, ...), Konjunktionen (und, oder, ...), ...
Anglizismen	Wortentlehnungen (z.B. Mail, Newsletter, Chat, Meeting, Update, ...)
Neologismen	Kunstwörter (z.B. Abmahnwelle, Nerd, googeln, verschlimmbessern, ...)
Wort n-Gramme	Ein kleines Textbeispiel $\xrightarrow{n=2}$ {(Ein kleines), (kleines Textbeispiel)}
Kollokationen	Häufig vorkommende Wortverbindungen (z.B. <i>starker Tobak</i> )
Wortarten	Adjektive, Interjektion, Numerale, Substantive, ...
Wortart n-Gramme	(Artikel-Adjektiv-Nomen), (Pronomen-Nomen-Artikel), ...
Phrasen/Redewendungen	Redensarten (z.B. <i>aus dem Nähkästchen plaudern</i> )
Satz-Anfänge/Endungen	Satzanfang(Nomen), Satzende(finites Verb), ...
Wort-Komplexität	Wörter bestimmter Länge, Wörter mit $x$ Vokalen
Satz-Komplexität	Sätze bestimmter Länge, Vorfeld/Mittelfeld/Nachfeld-Komplexitäten, ...



- Nach der Wahl geeigneter Feature-Kategorien, gilt es im nächsten Schritt Feature-Vektoren hinsichtlich der Texte zu generieren.
- Diese Vektoren dienen als Input für das jeweilige AV-Verfahren bzw. den zugrunde liegenden Machine Learning Algorithmen.
- Beispiel...

Unbekanntes  
Dokument



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

$$x_1, y_1 = \frac{\text{Anzahl aller "und"}}{\text{Anzahl aller Funktionswörter}}$$

Bekanntes  
Dokument



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

# Autorschaftsverifikation

- Um die Ähnlichkeit zwischen **X** und **Y** zu bestimmen, werden oftmals **Distanzfunktionen** eingesetzt.
- Eine davon ist z.B. die sogenannte

$$\text{ManhattanDistance}(X, Y) = \sum_{i=0}^n |x_i - y_i|$$

- Idee: Aufsummieren treppenartiger Abstände.
- Semantik: Je **kleiner** die **Distanz** ist, desto **ähnlicher** ist der **Schreibstil** !



[ManhDist]

- Beobachtung: Das Resultat einer Distanzfunktion ist ein Wert zwischen 0 und unendlich.
- Um ein Ähnlichkeitsvergleich durchzuführen bietet es sich an aus einer Distanzfunktion **dist(X, Y)** eine Ähnlichkeitsfunktion **sim(X, Y)** zu konstruieren:

$$\text{sim}(X, Y) = \frac{1}{1 + \text{dist}(X, Y)}$$

- Dadurch erhalten wir einen Wert in einem überschaubarem  $[0, 1]$  Intervall, mit dessen Hilfe eine Ähnlichkeit berechnet werden kann.

- Um zu erkennen **ab wann** eine Ähnlichkeit besteht, muss ein Schwellwert  $\theta$  aus den Trainingsdaten ermittelt werden.
- Wird dieser überschritten, so ist eine Ähnlichkeit gegeben und die **Autorschaft akzeptiert**.
- Um ein vernünftiges  $\theta$  zu erhalten muss das AV-Verfahren i.d.R. hinsichtlich ein zu definierendes Performanzmaß optimiert werden.
- Auf dem Gebiet der AV wurden bisher mehrere Performanzmaße vorgeschlagen...

# Performanzmaße

- Um die Güte von AV-Verfahren zu bestimmen sind Performanzmaße essentiell.
- Basis aller Performanzmaße sind die Ausprägungen der sogenannten **Konfusionsmatrix**:

		Predicted Class	
		$A$	$\neg A$
True Class	$A$	TP (True Positives / <i>hit</i> )	FN (False Negatives / <i>false alarm</i> )
	$\neg A$	FP (False Positives / <i>miss</i> )	TN (True Negatives / <i>correct rejection</i> )

- Ausgehend von dieser Matrix lassen sich zahlreiche Performanzmaße ableiten.

- Gängige Maße in der AV sind:

- $$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} = \frac{n_c}{n}$$

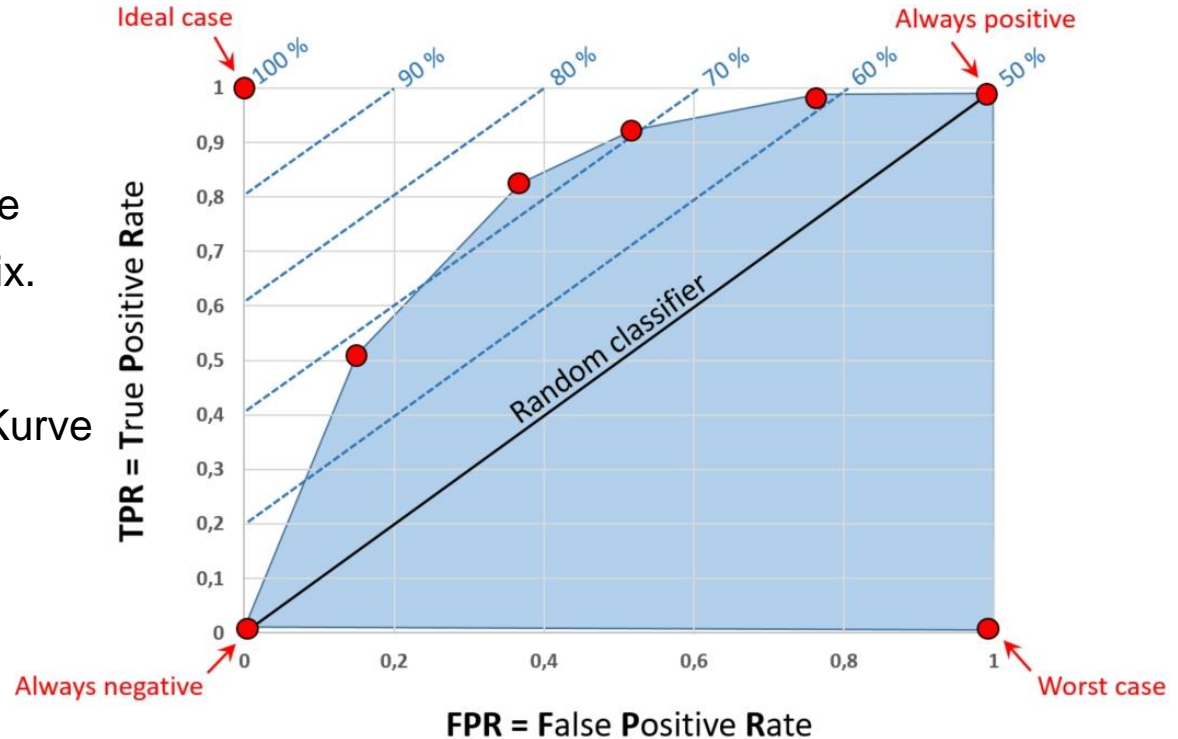
- $$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \text{ mit Precision} = \frac{n_c}{\text{Anzahl aller Antworten}}, \text{ Recall} = \frac{n_c}{n}$$

- $$c@1 = \frac{1}{n} \left( n_c + \left( \frac{n_u \cdot n_c}{n} \right) \right), \text{ mit } n_u = \text{Anzahl unbeantworteter Verifikationsfälle}$$



- **AUC (Area Under the Curve)**
- Jeder Punkt auf dem ROC-Kurve entspricht einer Konfusionsmatrix.
- Die Fläche unterhalb der ROC-Kurve ist die AUC...

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$



# COAV – Compression-based Authorship Verifier

- COAV ist ein von uns entwickeltes AV-Verfahren, welches in 2017 im Rahmen der ARES-Konferenz erstmals vorgestellt wurde.

**On the Usefulness of Compression Models  
for Authorship Verification**

Oren Halvani\*  
Christian Winter  
Lukas Graner  
Fraunhofer Institute for Secure Information Technology SIT  
Darmstadt, Germany  
<FirstName>.<Surname>@SIT.Fraunhofer.de

**ABSTRACT**  
Compression models represent an interesting approach for different classification tasks and have been used widely across many research fields. We adapt compression models to the field of authorship verification (AV), a branch of digital text forensics. The task in AV is to verify if a questioned document and a reference document

**ACM Reference format:**  
Oren Halvani, Christian Winter, and Lukas Graner. 2017. On the Usefulness of Compression Models for Authorship Verification. In *Proceedings of 12th International Conference on Availability, Reliability and Security, Reggio Calabria, Italy, August 29-September 01, 2017 (ARES 2017)*, 11 pages. <https://doi.org/10.1145/3098954.3104050>

[ARES2017]

- Im März 2018 folgte dann eine Verfeinerung des Verfahrens, die für forensische Zwecke prädestiniert ist, da hier keine Trainingsdaten benötigt werden (später mehr dazu).

- COAV verfolgt gegenüber existierender AV-Verfahren einen anderen Ansatz und delegiert den gesamten Feature-Engineering-Prozedere an einem **Kompressionsalgorithmus**.
- Dieser überführt die Texte in entsprechende Repräsentationen (Symbole mit Kontexten und deren Auftretenswahrscheinlichkeiten).
- Die Inhalte der komprimierten Daten können als individuelle **Modelle** aufgefasst werden die Texte kompakt beschreiben...

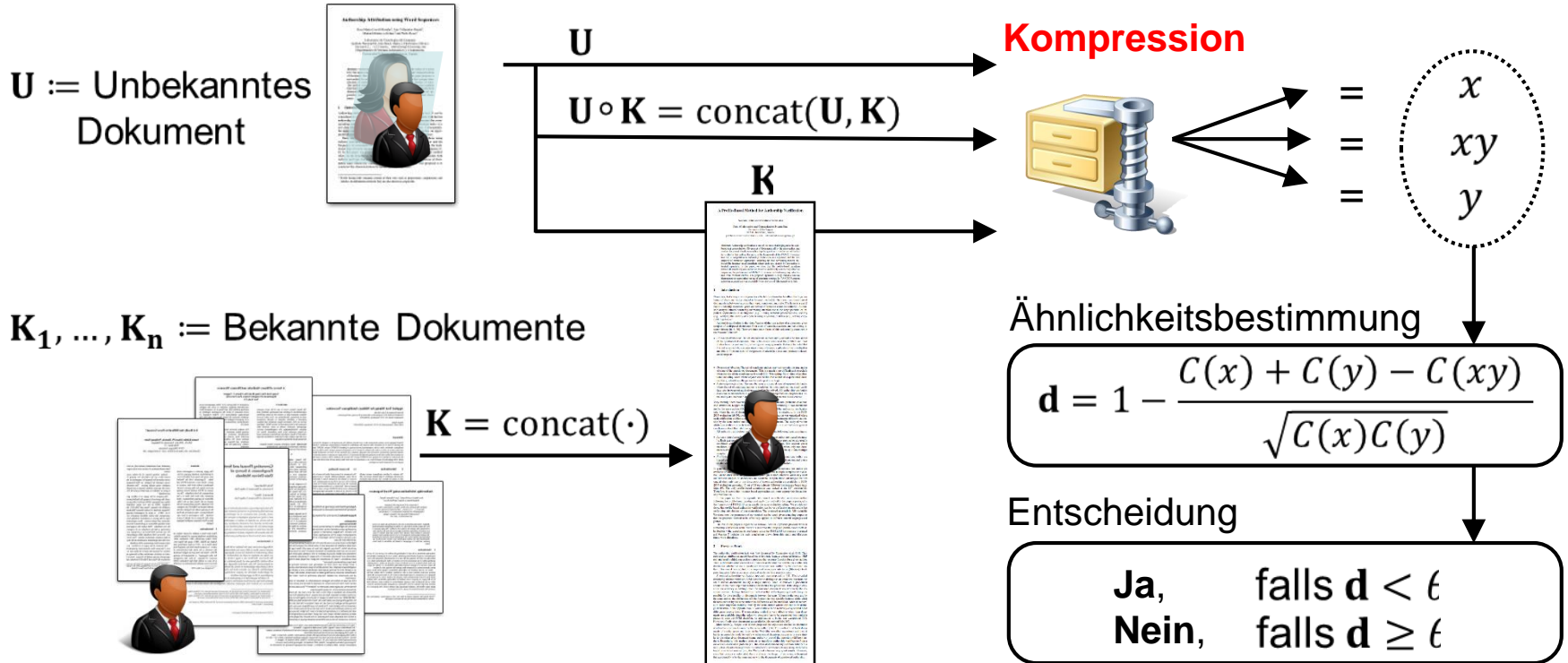
- Um die Texte in eine kompakte Repräsentation zu überführen nutzt COAV **statistische** Kompressoren (darunter **PPMd**, ZPAQ oder LZMA).
- Neben der Komprimierung der Texte wird zudem ein Maß hinsichtlich der **Ähnlichkeitsbestimmung** benötigt:

$$S(x, y) = 1 - \frac{C(x) + C(y) - C(xy)}{N(x, y)}$$

$C(\cdot)$  = Länge eines komprimierten Textes

$N(\cdot, \cdot)$  = Normalisierungsfaktorkomprimierten

- Der schematische Aufbau der Verfahrens sieht wie folgt aus...



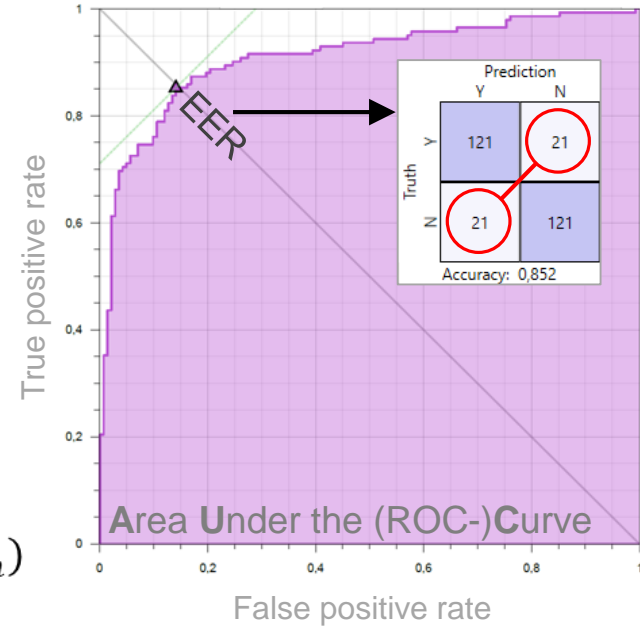
- Die Verifikationsentscheidung hängt (wie bei jedem anderen AV-Verfahren auch) von einem Schwellwert  $\theta$  ab. Um diesen zu ermitteln wird ein gleichverteilter Trainingskorpus benötigt.



- Für jedes Verifikationsproblem wird ein Ähnlichkeitswert  $d_i$  berechnet.

- Anschließend wird die Gleichfehlerrate (**Equal Error Rate**) für alle  $\mathbf{d}_1, \mathbf{d}_2 \dots, \mathbf{d}_n$  bestimmt. Die EER ist ein bekanntes Gütemaß, welches in der Biometrie oft verwendet wird.
- Visuell entspricht die EER demjenigen Punkt, an der die Hauptdiagonale die ROC-Kurve schneidet.
- Die Berechnung der EER erfolgt mittels:  $median(\mathbf{d}_1, \mathbf{d}_2 \dots, \mathbf{d}_n)$

$$\mathbf{Ja} = \begin{bmatrix} \mathbf{d}_1 & \mathbf{d}_2 & \mathbf{d}_3 \\ 0.0 & 0.1 & 0.9 \end{bmatrix} \quad \mathbf{Nein} = \begin{bmatrix} \mathbf{d}_4 & \mathbf{d}_5 & \mathbf{d}_6 \\ 0.1 & 0.2 & 0.3 \end{bmatrix} \quad \longrightarrow \theta = 0.15$$





- COAV wurde auf dem Testkorpus des internationalen Wettbewerbs **PAN-2015** evaluiert und erzielte ähnliche Ergebnisse zu denen des Gewinners...
- ...allerdings brauchte es nur 7 Sekunden statt knapp 22 Stunden, um 500 Verifikationsfälle zu bearbeiten.
- COAV schlug unser bisheriges Verfahren "AVeer" aus dem Jahre 2016.



[DFRWS2016]

Team	FS	AUC	c@1	UP	Runtime
Bagnall	0.614	0.811	0.757	3	21:44:03
<b>Our approach</b>	<b>0.605</b>	<b>0.802</b>	<b>0.754</b>	<b>0</b>	<b>00:00:07</b>
Castro-Castro et al.	0.520	0.750	0.694	0	02:07:20
Gutierrez et al.	0.513	0.739	0.694	39	00:37:06
Kocher and Savoy	0.508	0.738	0.689	94	00:00:24
PAN15-ENSEMBLE	0.468	0.786	0.596	0	-
Halvani	0.458	0.762	0.601	25	00:00:21
Moreau et al.	0.453	0.709	0.638	0	24:39:22
Pacheco et al.	0.438	0.763	0.574	2	00:15:01
Hürlimann et al.	0.412	0.648	0.636	5	00:01:46
PAN14-BASELINE-2	0.409	0.639	0.640	0	00:26:19
PAN13-BASELINE	0.404	0.654	0.618	0	00:02:44
Posadas-Durán et al.	0.400	0.680	0.588	0	01:41:50
Maitra et al.	0.347	0.602	0.577	10	15:19:13
Bartoli et al.	0.323	0.578	0.559	3	00:20:33
Gómez-Adorno et al.	0.281	0.530	0.530	0	07:36:58
Solórzano et al.	0.259	0.517	0.500	0	00:29:48
Nikolov et al.	0.258	0.493	0.524	16	00:01:36
Pimas et al.	0.257	0.507	0.506	0	00:07:22
PAN14-BASELINE-1	0.249	0.537	0.464	159	00:01:11
Mechti et al.	0.247	0.489	0.506	0	00:04:59
Sari and Stevenson	0.201	0.401	0.500	0	00:05:47
Vartapetian and G.	0.000	0.500	0.000	500	-

- Neben dem PAN-2015 Wettbewerb wurde COAV auf andere Korpora mit 7 weiteren Sprachen (außer nur Englisch) evaluiert, mit sprachübergreifend stabilen Ergebnissen.
- Hauptvorteil des Verfahrens: **Keine explizite Definition von Features !**  
Stattdessen delegiert das Verfahren diesen Prozess an einem Kompressionsalgorithmus.
- Einzige Abhängigkeit: Ein Schwellwert muss erlernt werden, um die Verifikation zu ermöglichen.
- Lässt sich diese **Abhängigkeit** ebenfalls **eliminieren** ?

CHALLENGE ACCEPTED



# OCCAV – One-Class Compression-based Authorship Verifier

- OCCA ist unsere jüngste AV-Methode, die 2018 auf der ECIR-Konferenz vorgestellt wurde  
→ <https://www.youtube.com/watch?v=WqxlVzTrUZg>

## Authorship Verification in the Absence of Explicit Features and Thresholds

Oren Halvani<sup>(✉)</sup>, Lukas Graner, and Inna Vogel

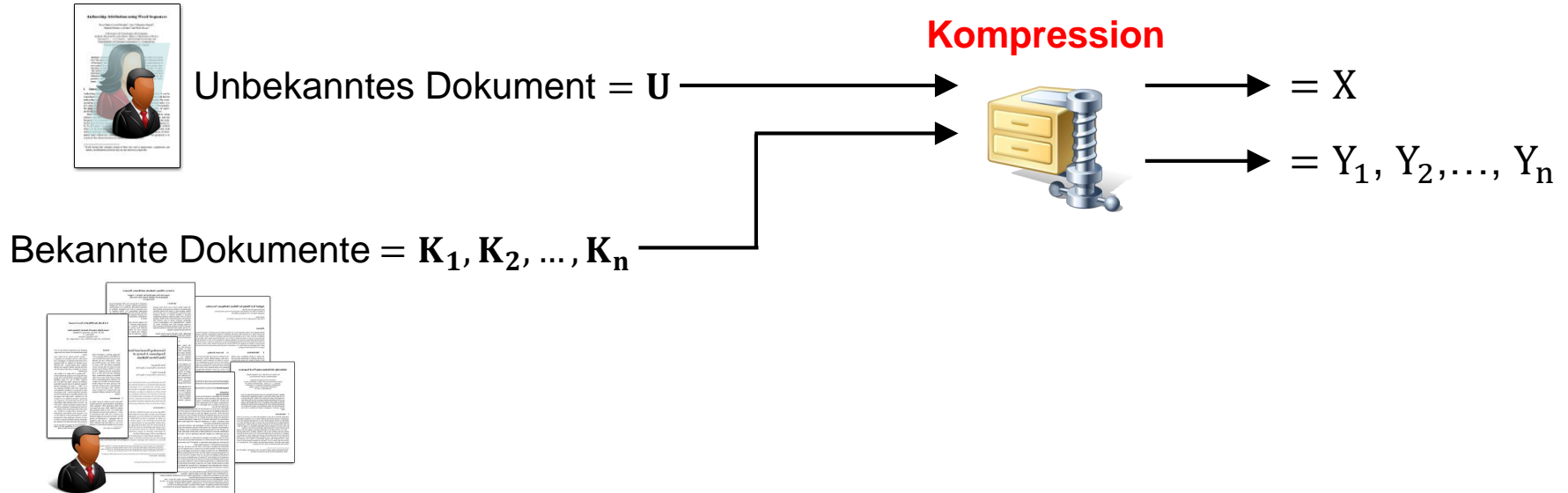
Fraunhofer Institute for Secure Information Technology,  
Rheinstraße 75, 64295 Darmstadt, Germany  
[Oren.Halvani@SIT.Fraunhofer.de](mailto:Oren.Halvani@SIT.Fraunhofer.de)

**Abstract.** Enhancing information retrieval systems with the ability to take the writing style of people into account opens the door for a number of applications. For example, one can link articles by authorships that can help identifying authors who generate hoaxes and deliberate misinformation in news stories, distributed across different platforms. Author-

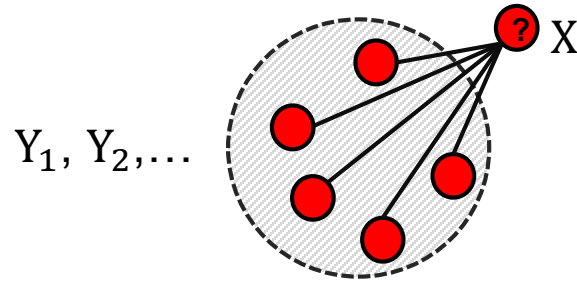
[ECIR2018]

- Das Verfahren ist im Wesentlichen eine kleine Modifikation von COAV.
- Es werden die selben Komponenten verwendet, allerdings auf eine andere Art und Weise.
- Wesentliche Eigenschaft von OCCAV ist die Tatsache das es "**out-of-the-box**" verwendet werden kann.
- Die typische Lernphase (Training) die AV-Verfahren i.d.R. benötigen entfällt dadurch gänzlich.

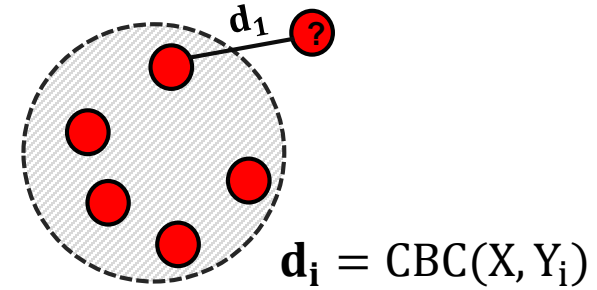
- **Änderung:** Statt alle  $K_i$ 's zu konkatenieren werden diese, so wie sie sind, komprimiert.



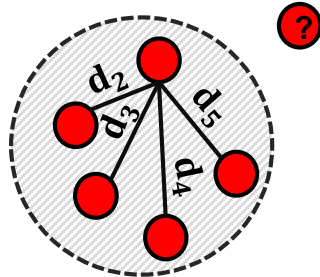
1.) Äußere Ähnlichkeiten berechnen



2.) Kleinste Ähnlichkeit notieren



3.) Innere Ähnlichkeiten berechnen



4.) Entscheidung:

Autorschaft akzeptiert, falls  $d_1 < \frac{1}{n} \sum_{i=2}^n d_i$   
ansonsten verworfen.

- OCCAV wurde auf 8 Korpora (1 Korpus = 1 Sprache) gegenüber 6 Baselines evaluiert...
- ...und erzielte sprachübergreifend hohe und stabile Ergebnisse, auf Basis von **AUC**.

Corpus ID	Language	OCCAV	Baselines						avg (·)
			COAV	CNG	Prof-AV	GLAD	GI	Impr. GI	
L'Express	French	0,841	<b>0,940</b>	0,698	0,735	0,799	0,929	0,897	0,834
Trouw	Dutch	0,867	<b>0,920</b>	0,745	0,827	0,806	0,831	0,880	0,839
Dziennik	Polish	0,937	<b>0,968</b>	0,756	0,916	0,854	0,945	0,938	<b>0,902</b>
El Pais	Spanish	0,908	<b>0,956</b>	0,866	0,881	0,907	0,937	0,934	<b>0,913</b>
Svenska	Swedish	<b>0,909</b>	0,782	0,681	0,742	0,194	0,769	0,856	0,705
Enron	English	0,823	0,878	0,777	0,840	0,849	<b>0,943</b>	0,923	0,862
Gutenberg	German	0,821	0,903	0,784	0,821	0,863	0,909	<b>0,932</b>	<b>0,862</b>
Cookpad	Greek	<b>0,878</b>	0,806	0,783	0,761	0,774	0,839	0,748	0,799
avg (·)		0,873	<b>0,894</b>	0,761	0,815	0,756	<b>0,888</b>	<b>0,889</b>	



- OCCAV lernt **individuelle Schwellwerte** direkt aus den Daten, sodass kein globaler Schwellwert definiert werden muss.
- Da das Verfahren (analog zu COAV) keine expliziten Features benötigt, ist es losgelöst von einer entsprechenden Trainingsprozedur.
- Diese Eigenschaft ist insbesondere wichtig in Szenarien, in denen keine Trainingsdaten existieren oder deren Beschaffung zu aufwändig / teuer ist.
- OCCAV wurde im Rahmen des BMBF-Projekts EWW zur Bekämpfung von Versicherungsbetrug erfolgreich eingesetzt.

Vielen Dank für die Aufmerksamkeit !

...Fragen?



[Oren.Halvani@SIT.Fraunhofer.de](mailto:Oren.Halvani@SIT.Fraunhofer.de)

- [ManhDist] Alexey Grigorev. *What is the difference between Manhattan and Euclidean distance measures?* <https://www.quora.com/What-is-the-difference-between-Manhattan-and-Euclidean-distance-measures>
- [DFRWS2016] Oren Halvani, Christian Winter, Anika Pflug. *Authorship verification for different languages, genres and topics.* <https://www.sciencedirect.com/science/article/pii/S1742287616000074>
- [ARES2017] Oren Halvani, Christian Winter, Lukas Graner. *On the Usefulness of Compression Models for Authorship Verification.* <https://dl.acm.org/citation.cfm?id=3104050>
- [ECIR2018] Oren Halvani, Lukas Graner, Inna Vogel. *Authorship Verification in the Absence of Explicit Features and Thresholds.* [https://link.springer.com/chapter/10.1007%2F978-3-319-76941-7\\_34](https://link.springer.com/chapter/10.1007%2F978-3-319-76941-7_34)