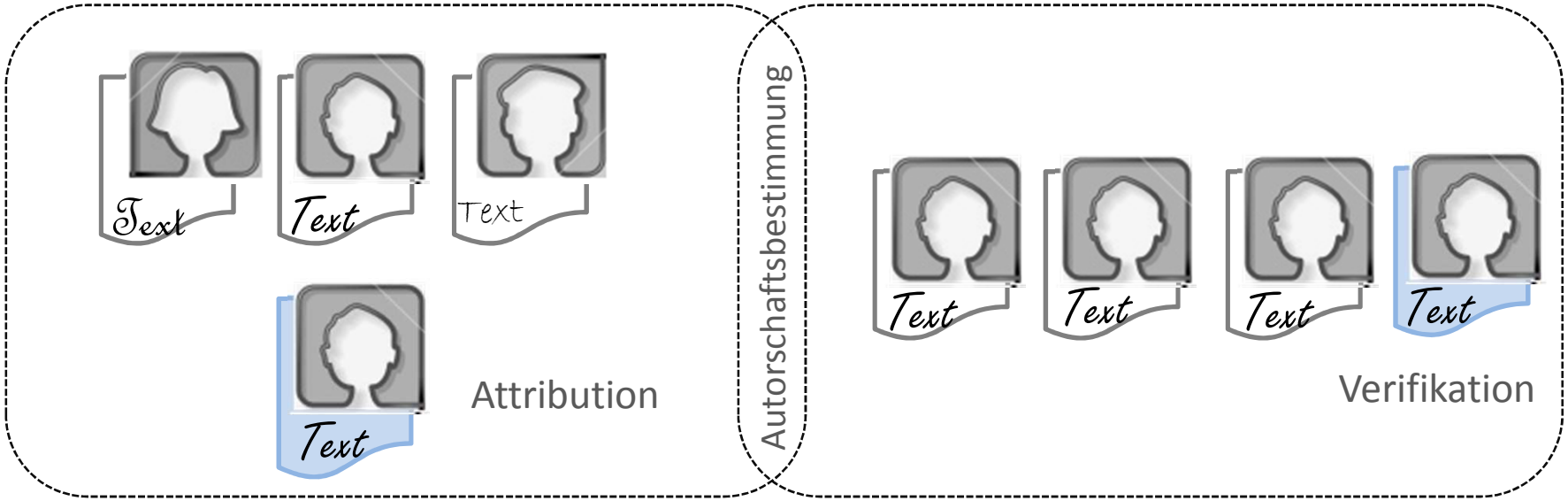
Character-N-Gramm-basierte Retrieval-Modelle in der Autorschaftsbestimmung

Michael Blersch

Autorschaftsbestimmung

- Verfahren die auf Basis des Schreibstils eines Textes rückwirkend auf die Autorschaft schließen.
- Der Schreibstil eines Autors wird durch Merkmale (Stilmerkmale) aus einem Text bestimmt.
- Verfahren zur Lösung behandeln das Attributions- oder das Verifikationsproblem von Autoren.

Problemklassen und Anwendungsfälle



- Streit über die Urheberschaft eines Textes.
- Identifikation des Verfassers eines Drohbriefes oder von belästigenden Nachrichten.
- Verifikation der Authentizität von Abschiedsbriefen.

Analyseverfahren zur Autorschaftsbestimmung:

Anforderungen an Stilmerkmale zur Erfassung und Abgrenzung von Schreibstilen:

- Hohes Unterscheidungspotential bei unterschiedlichen Autoren.
- Hohe Toleranz bei Stilvariationen gleicher Autoren.
- Keine Messung von inhaltlicher Information.

Herausforderungen für Analyseverfahren:

- Anzahl der in Frage kommenden Autoren.
- Limitierte Texte (Textlänge, Textmengen pro Autor).
- Umgang mit dem sog. „open candidate set “.

Beiträge

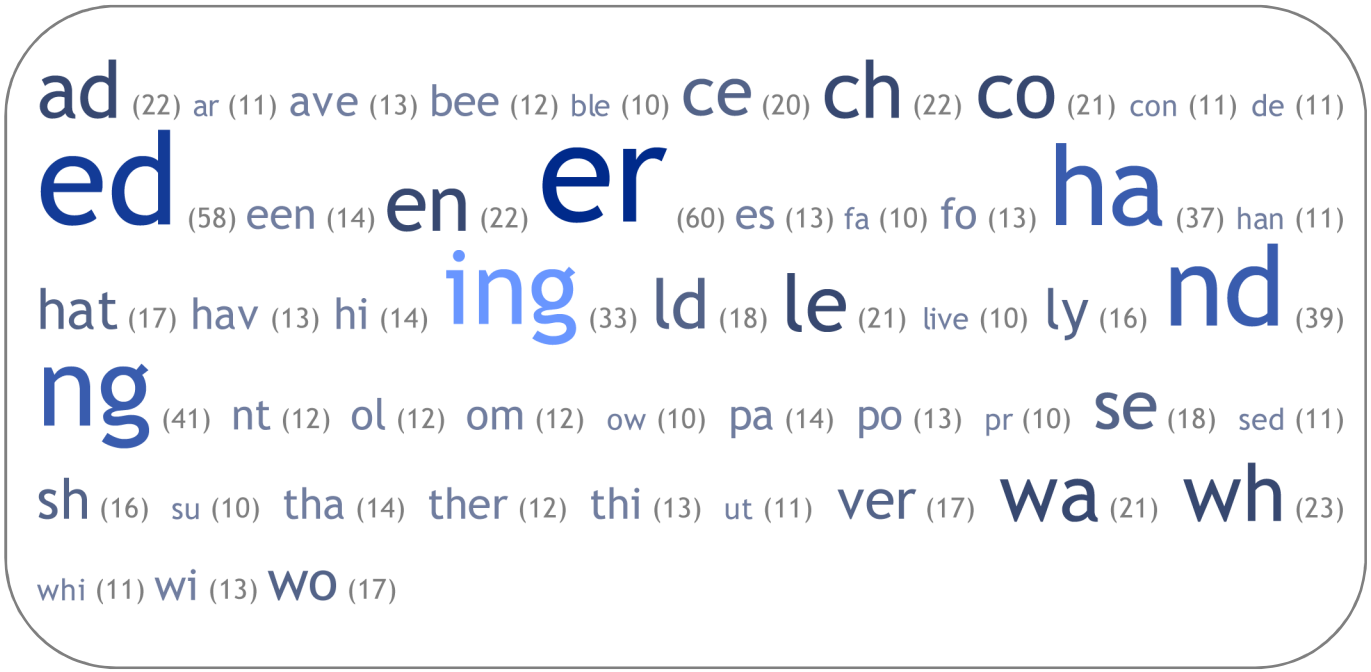
Character-n-Gramm basierte Modelle zur Lösung des Attributionsproblems von Autoren:

- Einführung des ESA-Modells
  - Gegenüberstellung der Stilrepräsentation des ESA-Modells mit dem Vektorraummodell.
  - Evaluierung der Stilrepräsentationen beider Modelle durch Klassifikatoren.

- Analyse des Projektionsmodells von Koppel et. Al.
  - Vorschläge zur Verbesserung der Entscheidungsfunktion des Modells.
  - Evaluierung der Modifikationen des Modell.

Stilmerkmale – Character-n-Gramme

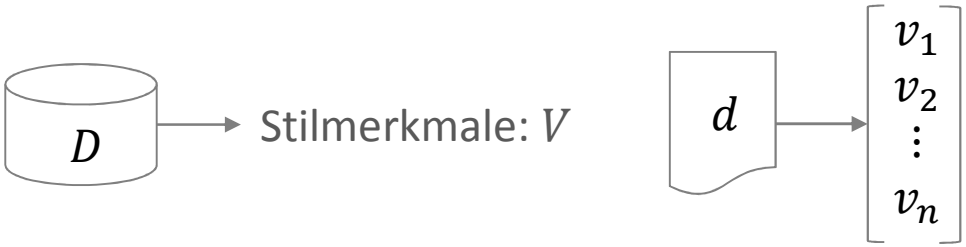
- Text wird als eine Sequenz von Zeichen betrachtet.
- Positionsweise werden n-Zeichen aus einem Text extrahiert.



Text: Oliver Twist Kapitel 1 (1000 Wörter), Autor Charles Dickens.
Claud aus 50 von ca. 1800 Characer-3-Grammen.

Vektorraummodell (VSM)

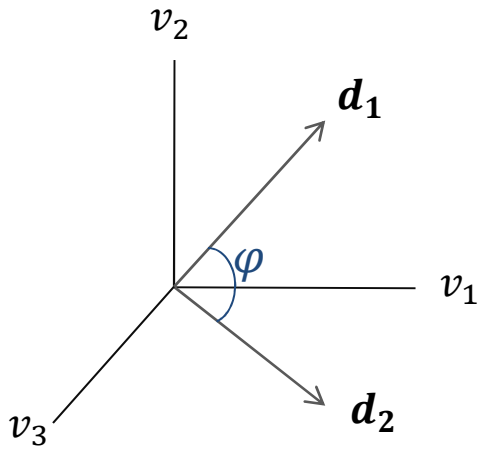
- Stilrepräsentation eines Dokuments d:



- Gewichtung: $tf(v,d) \cdot idf(v,D) = tf \cdot \log_2\left(\frac{|D|+1}{df(v,D)+1}\right)$

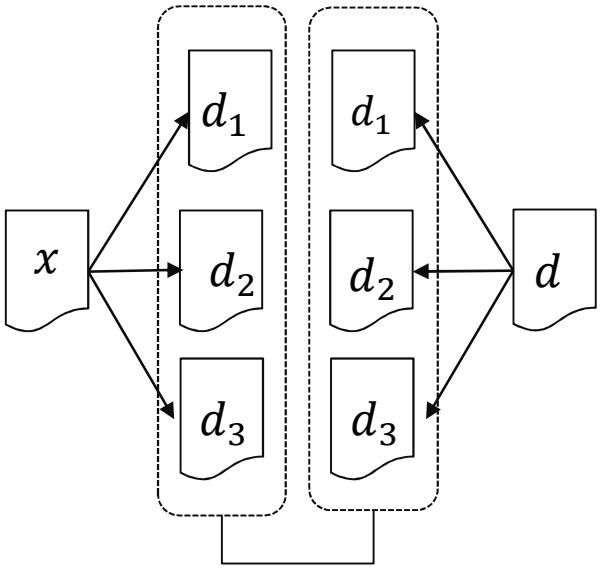
- Stilähnlichkeit zwischen Dokument d_1 und Dokument d_2:

$$\rho_{VSM} = \varphi_{\cos\,(d_1,d_2)} = \mathbf{d}_1^{\mathrm{T}} \cdot \mathbf{d}_2 \text{ , mit } \|\boldsymbol{d}_1\| = \|\boldsymbol{d}_2\| = 1$$

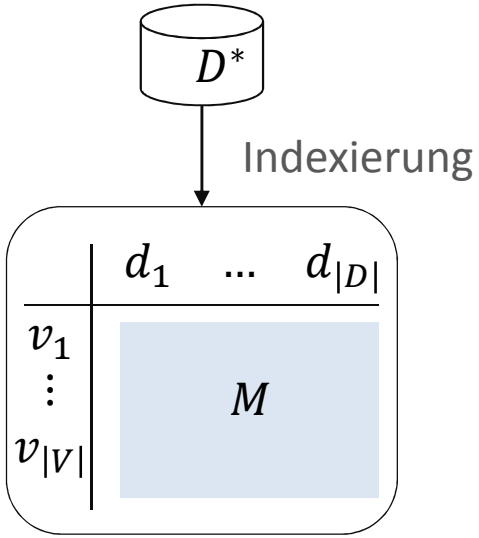


Explicit Semantic Analysis (ESA)

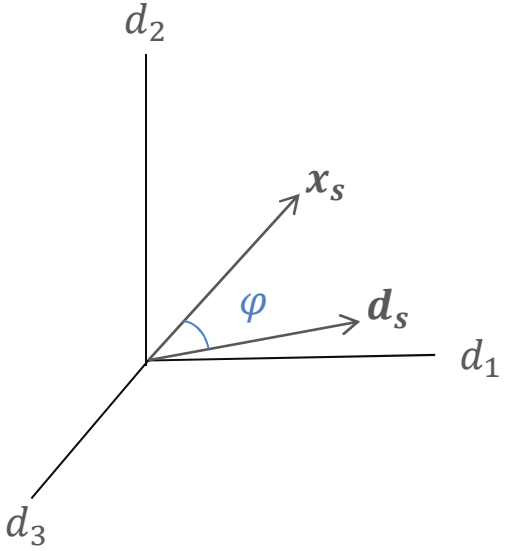
- Verwendung von externen Dokumenten D^*.
- Repräsentation von d_1 und d_2 als Stilähnlichkeitsvektoren $\boldsymbol{x}_s$ und $\boldsymbol{d}_s$.



$d_i \in D^*$



$\boldsymbol{x}_s = M^T \cdot \boldsymbol{x}$

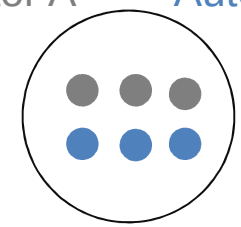


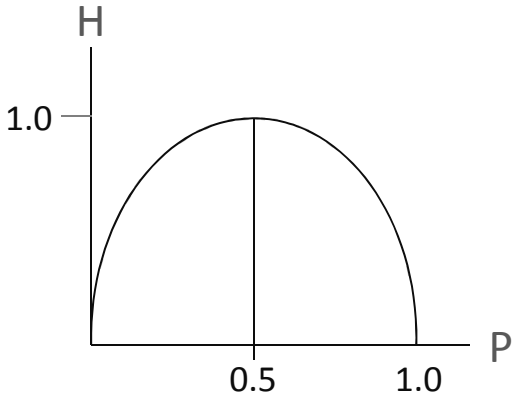
$\rho_{ESA} = \boldsymbol{x}_s \cdot \boldsymbol{d}_s$

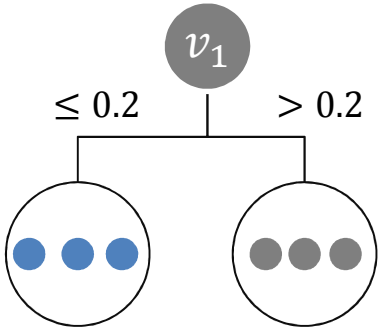
Merkmalsselektion: Entropie und information gain

- Entropie H zur Verteilung der Autoren C:

$$H(C) = \sum_{i=1}^{|C|} P(c = c_i) \cdot \log_2 P(C = c_i)$$





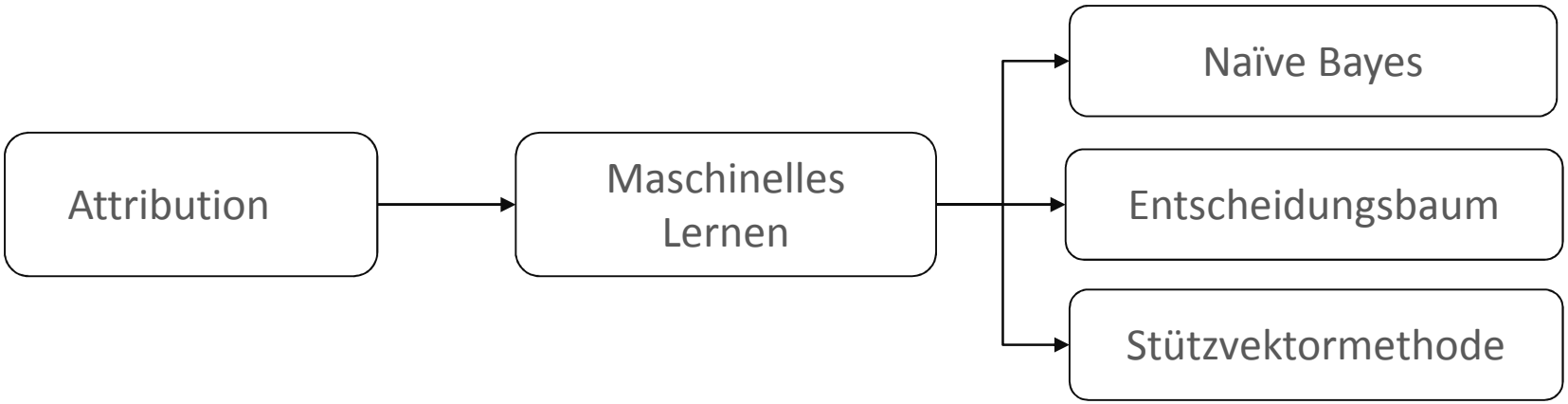


- Information gain als Maß für die Reduktion der Unreinheit durch ein Stilmerkmal v_k:

$$\text{gain} = H(C) - H(C|v_k) \qquad H(C|v_k) = \sum_{j=1}^{|k|} P\left(v_k = v_{kj}\right) \cdot H\left(C\middle|v = v_{kj}\right)$$

Überwachtes Lernen

- Textrepräsentation des Schreibstils in Form eines Vektors.
- Klassifikationsalgorithmus lernt auf Textbeispielen von bekannten Autoren ein Attributionsmodell.
- Klassifikation eines anonymen Textes erfolgt durch das gelernte Modell.

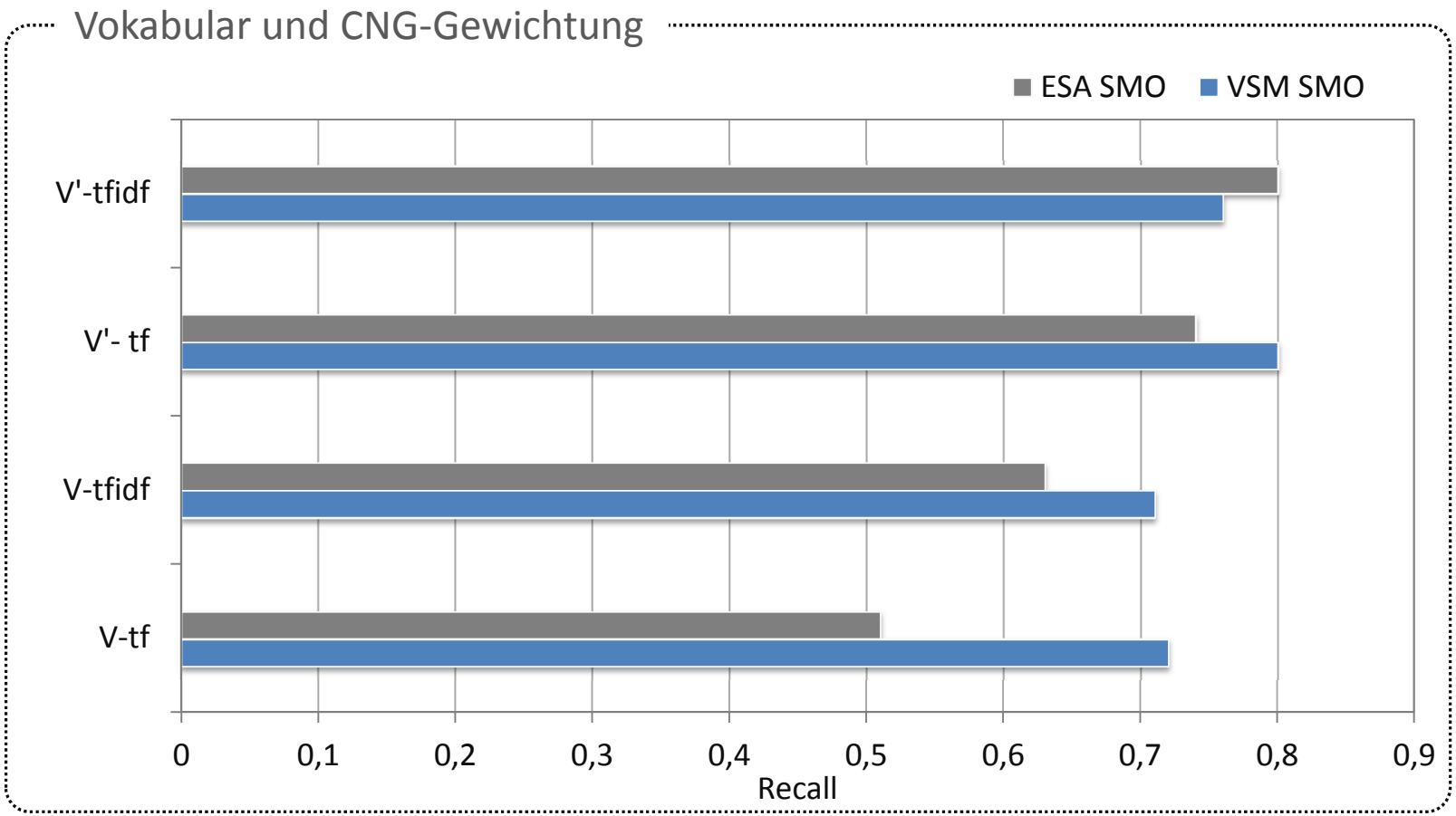


Evaluierung der Stilrepräsenationen: ESA versus VSM

- Als Stilmerkmale werden Character-3-Gramme verwendet.
- Der Korpus enthält 9 Autoren mit je 2 Bücher [nach Koppel].
- Trainingsmenge: Zerlegung des jeweils ersten Buches je Autor in Dokumente mit 500 Worten.
- Testmenge: Zerlegung des jeweils zweiten Buches je Autor in Dokumente mit 1.500 Worten.

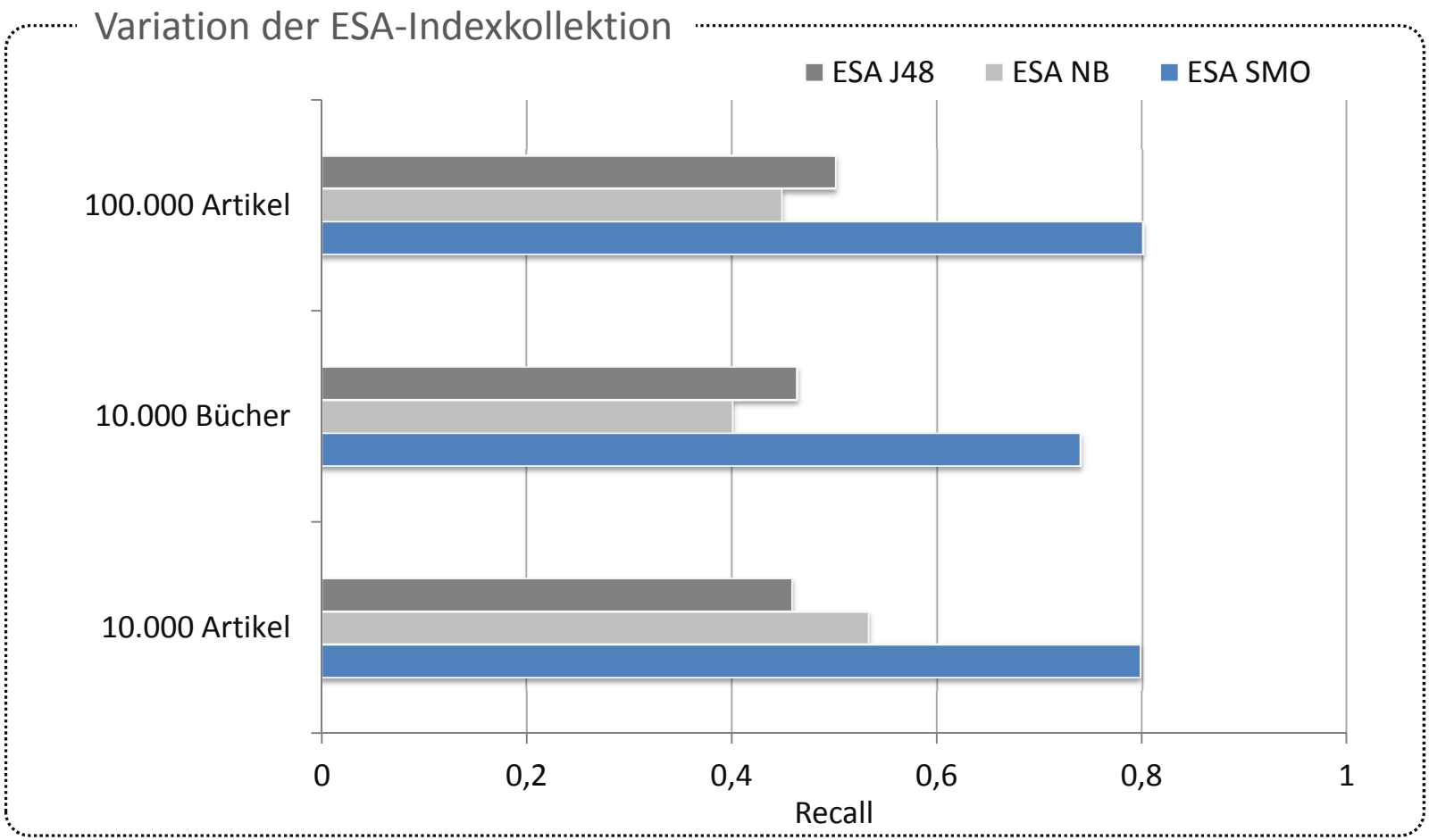
Modell	Representation	Klassifikator
ESA	Gewichtungen: tf / tf idf	Naive Bayes: NB
VSM	Auswahlmenge der Merkmale	Entscheidungsbaum: J48
		Stützvektormaschine: SMO

Evaluierung der Stilrepräsenationen: ESA versus VSM

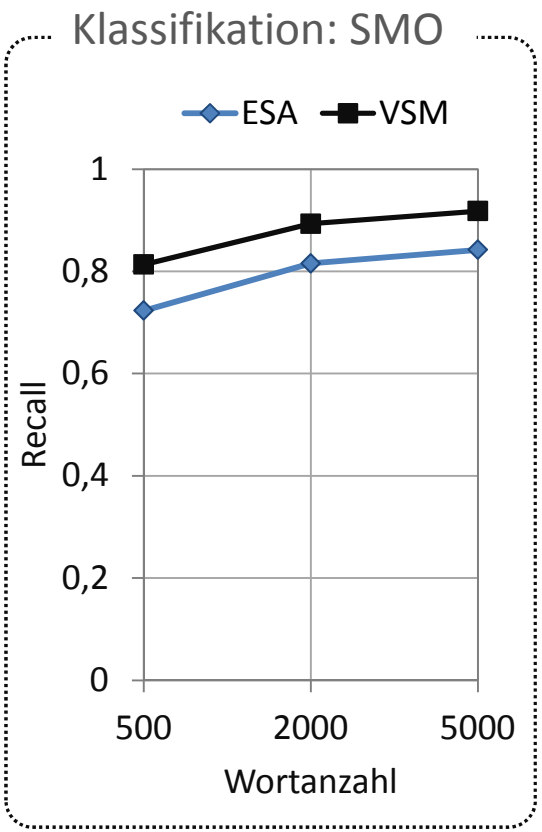


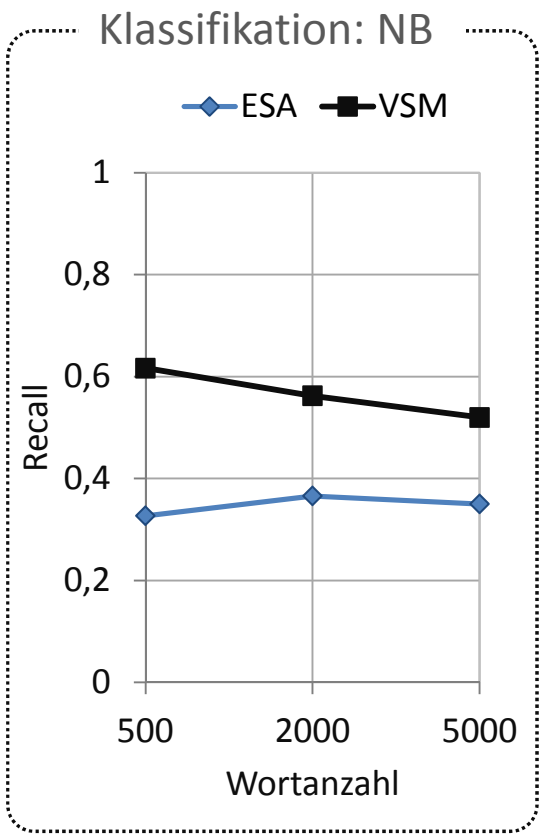
- Das Vokabular V enthält die 10.000 häufigsten C3G aus der Trainingsmenge.
- Das Vokabular V' enthält 1.000 C3G aus V mit dem höchsten information gain.
- Die ESA-Indexkollektion enthält 10.000 Wikipedia-Artikeln.

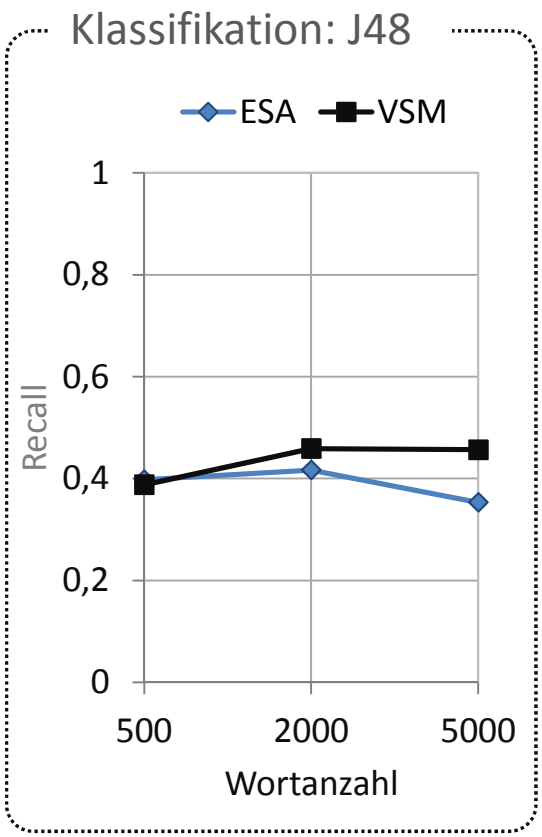
Evaluierung der Stilrepräsenationen: ESA versus VSM



Evaluierung der Stilrepräsenationen: ESA versus VSM







Projektionsmodell nach Koppel et. al.

Problemstellung:

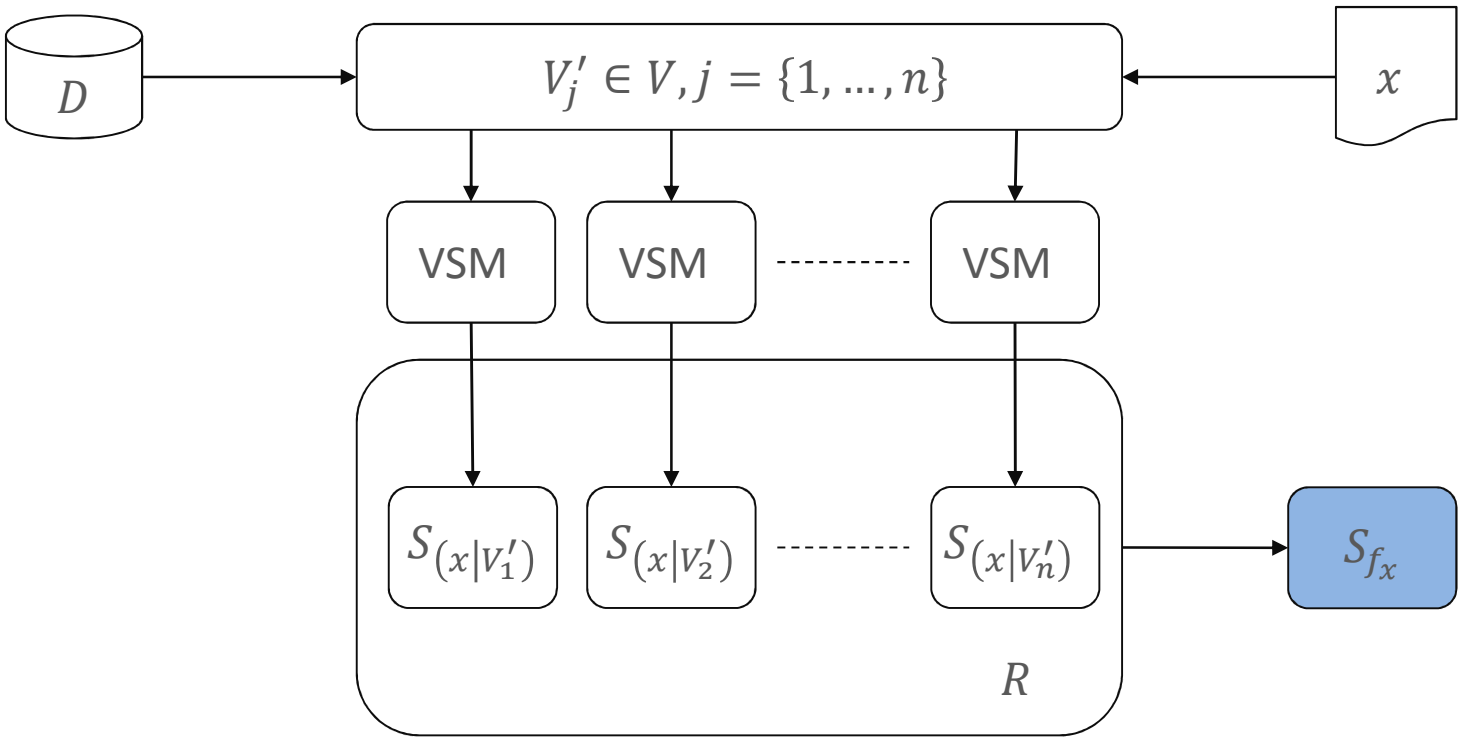
- Für die Autorschaft eines Dokuments können tausende potentielle Autoren in Frage kommen.
- Maschinelle Lernverfahren bieten keine adäquate Lösung aufgrund der hohen Autorenanzahl.

Ansatz des Modell:

- Um Schreibstile zu vergleichen wird das Vektorraummodell verwendet.
- Für ein „anonymes“ Dokument x gibt es genau ein Ranking zu den „bekannten“ Dokumenten D.
- Koppel schlägt vor mehrere Rankings zu berechnen indem die Stilrepräsentation variiert wird.
- Die Rankings bilden die Grundlage für die Entscheidungsfunktion.

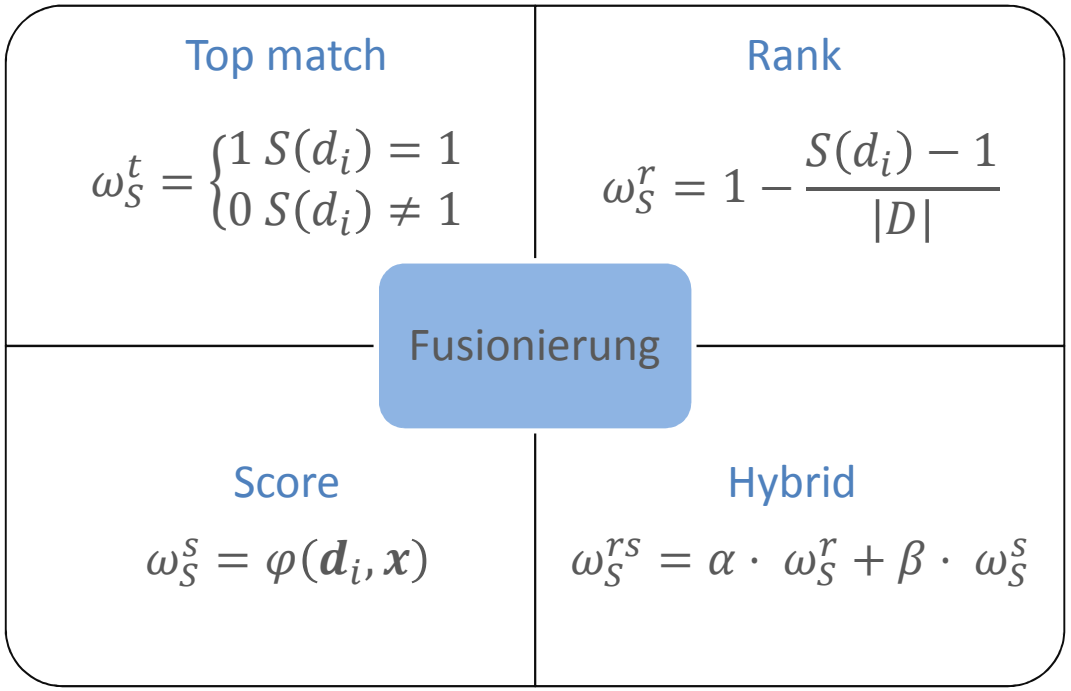
Projektionsmodell nach Koppel et. al.

- S ist ein Ranking von x über D.
- Die Menge R enthält n Rankings.
- V ist eine Menge an Stilmerkmalen und V' eine zufällige Auswahl mit $|V'| = \text{const}$



Projektionsmodell: Ranking-Fusionierungsmethoden

- Fusionswert eines Dokuments über alle Rankings: $f_x(d_i) = \sum_{S \in R} \omega_S(d_i)$



- Entscheidungsfunktion mit Sicherheitsschwelle σ:

- $\text{Autor von } x = \begin{cases} \text{Autor von } d_i \text{ mit } f_x(d_i) = S_{f_x}(1), \text{wenn } f_x(d_i) > \sigma \\ \text{unbekannt, sonst} \end{cases}$

Evaluierung des Projektionsmodells

- Korpus: 7.085 Autoren mit je einem Buch [Project-Gutenberg].
- Zerlegung der Bücher in Dokument x mit 500 und Dokument d mit 1.500 Worten.
- Die Menge X enthält 1.000 anonymen Dokumente x.
- Die Menge D enthält alle bekannten Dokumente d.

Stilrepräsentation

- C3G, $V = 95.000$
- C4G, $V = 500.000$
- C4G, $V = 100.000$
- C5G, $V = 100.000$
- $|V'| = 1.000$

Fusionierung

- Score based
- Rank based
- Hyprid

Ergebnis

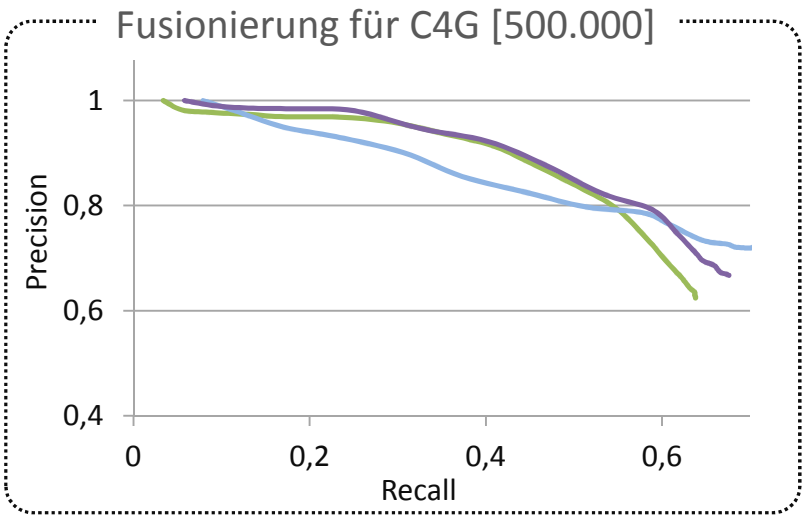
Recall-Precision-Graph:

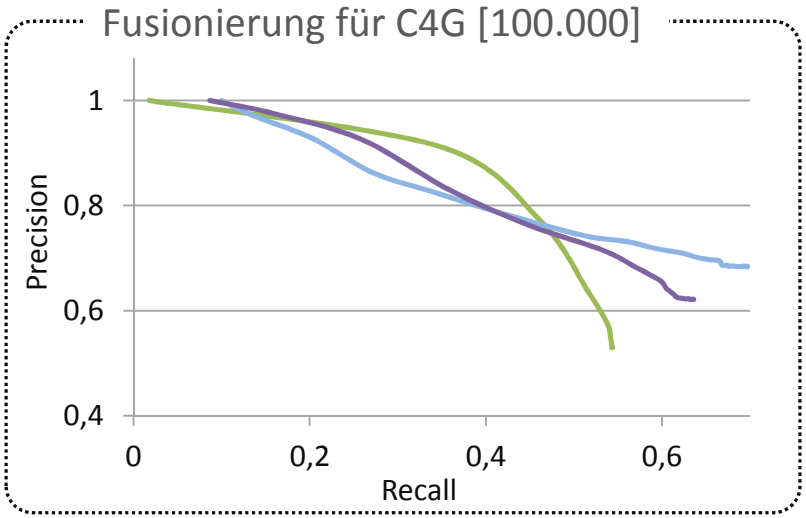
$|R| = n = 100$

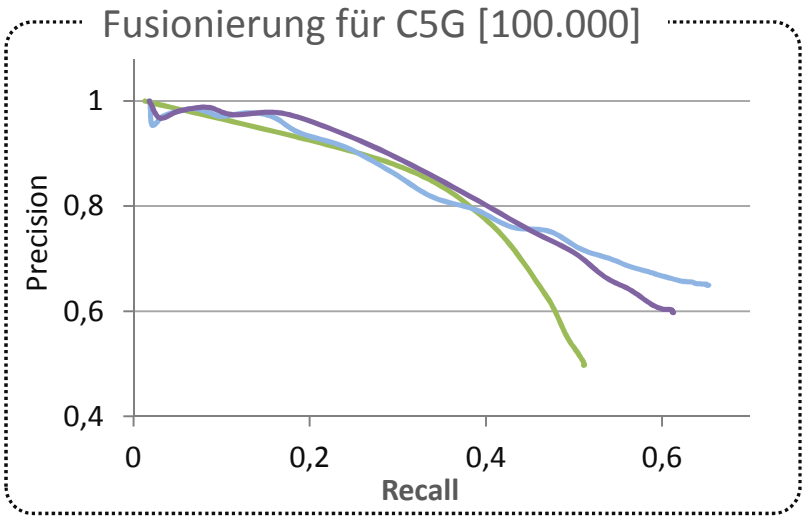
$\sigma = \{1, \dots, 100\}$

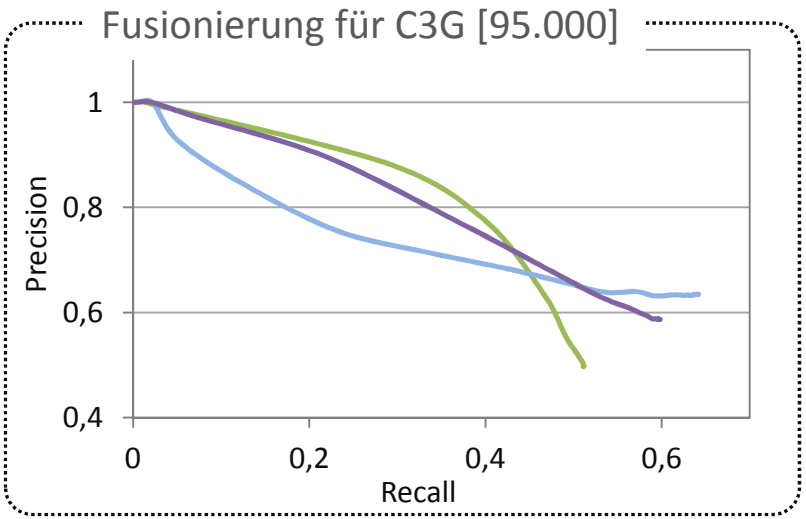
Evaluierung - Fusionierungsmethoden

Score Rank Hybrid









Zusammenfassung

Stilrepräsentationen im Modellvergleich – ESA-Modell versus Vektorraummodell.

- Vorselektion von Stilmerkmalen mittels information gain ist für beide Modell sinnvoll.
- Das ESA-Modell ist in der Lage Autoren zu unterscheiden, dabei sind Stilähnlichkeitsvektoren keine Verbesserung zur Stilrepräsentation im VSM.
- Im besten Fall unterscheiden beide Modell gleich gut.

Zusammenfassung

Evaluierung des Projektionsmodells:

- Zur Unterscheidung von tausenden Autoren sind Character-4-Gramme gut geeignet.
- Die Rangfusionierung erzeugt einen guten tradeoff zwischen Precision und Recall unter Berücksichtigung der Autorenanzahl und erzieht die größte Sicherheit bei einer Entscheidung.