Paarweise Autorenschaftsverifikation von kurzen Texten

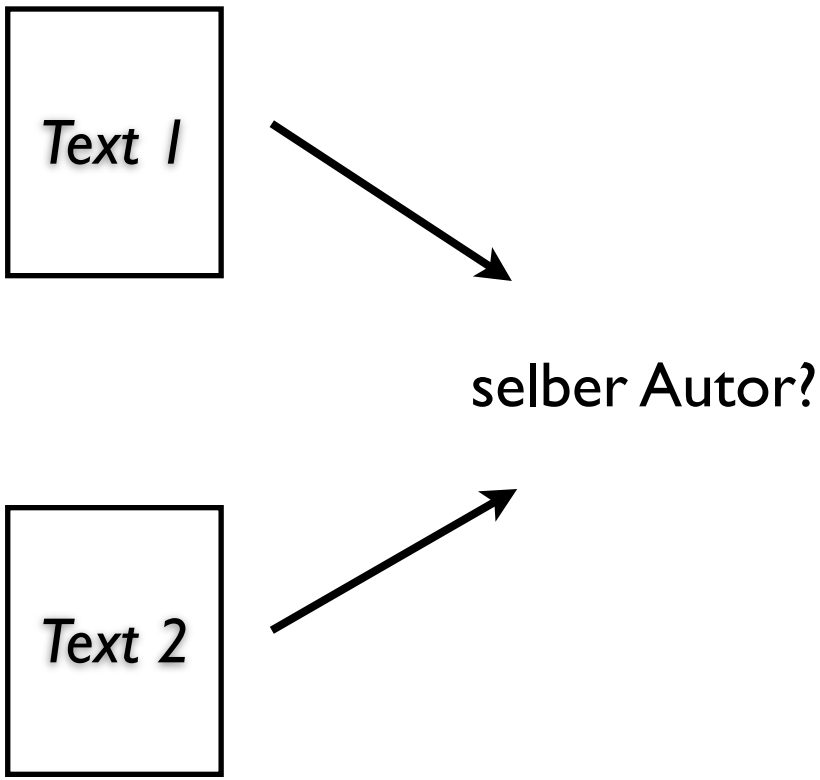
Verteidigung der Masterarbeit

Fabian Loose

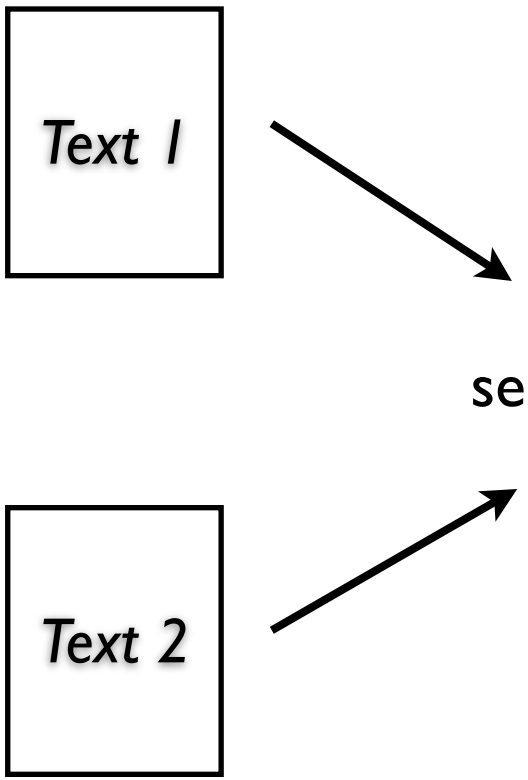
Gliederung

1. Was ist Autorenschaftsverifikation
2. Aktueller State-of-the-Art Algorithmus
3. Neu entwickelter Algorithmus
4. Evaluierung
5. Zusammenfassung

Problemstellung



Problemstellung



- intrinsische Plagiatanalyse
- Forensik
- Literaturwissenschaften

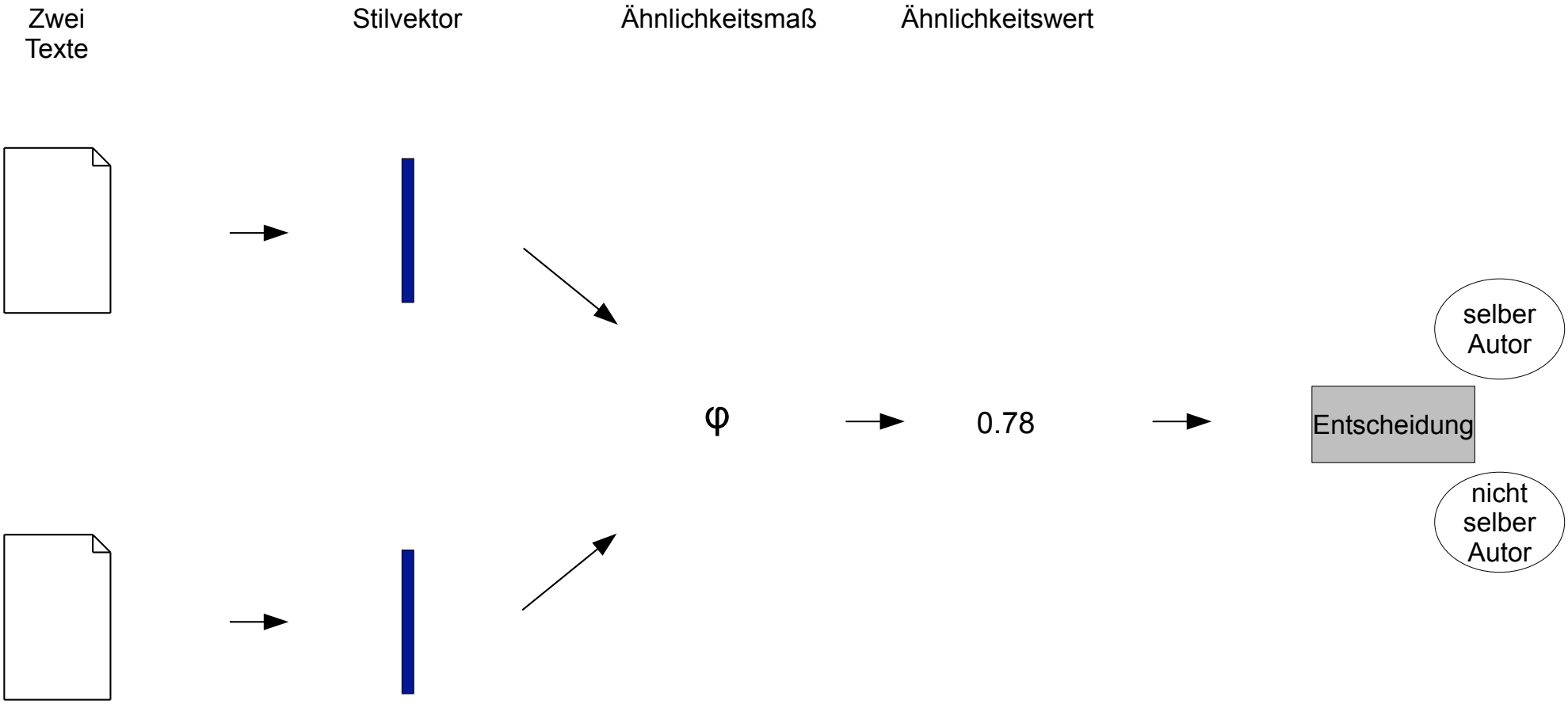
Stilmerkmale

- Relative Häufigkeiten von Elementen im Text

“Das ist ein Beispiel.”

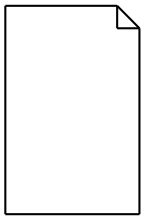
- Bekannte Stilmerkmale
  - Stoppwörter: *das, ist, ein*
  - Zeichen-Trigramme: *das, as_, s_i, _is, ist, st_, t_e, ... , pie, iel*
- Neu entwickelte Stilmerkmale
  - Wortlängen-Bigramme: *3-3, 3-3, 3-8*
  - Suffix-Bigramme: *as-st, st-in, in-el*

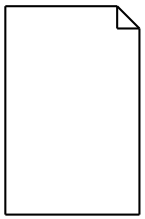
Lösungsschema



Unmasking [Koppel 2004]

Zwei
Texte

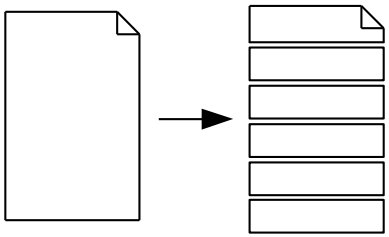


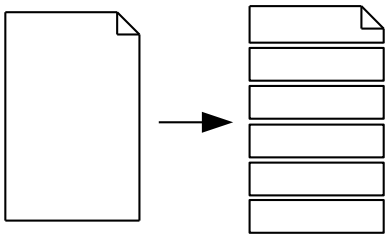


Unmasking

Zwei Texte

Chunking, 500 Wörter pro Chunk



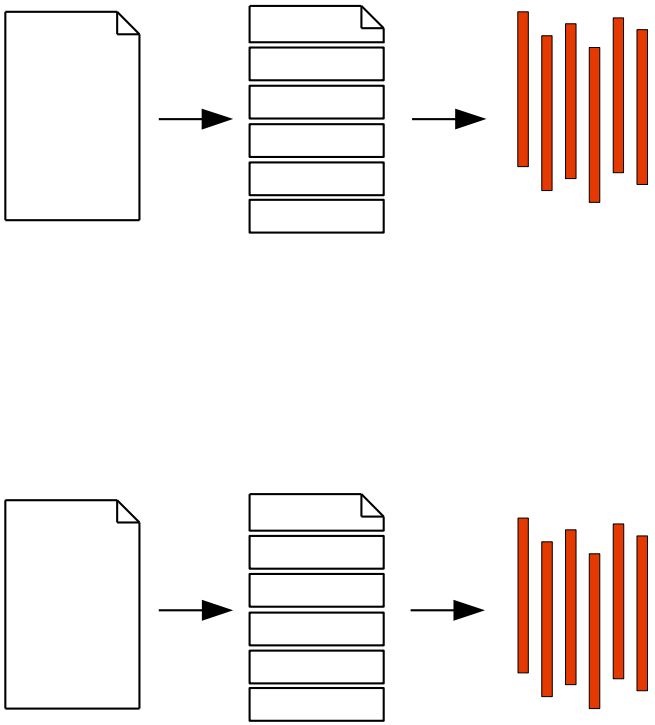


Unmasking

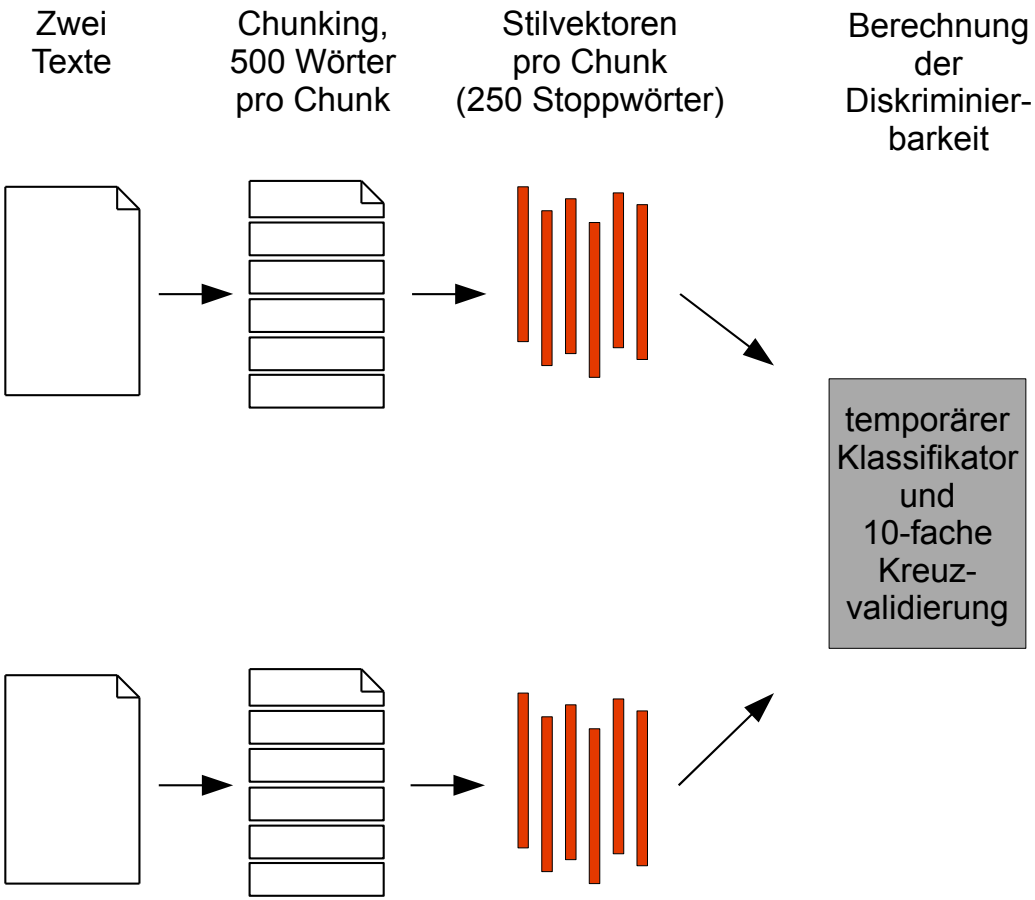
Zwei Texte

Chunking, 500 Wörter pro Chunk

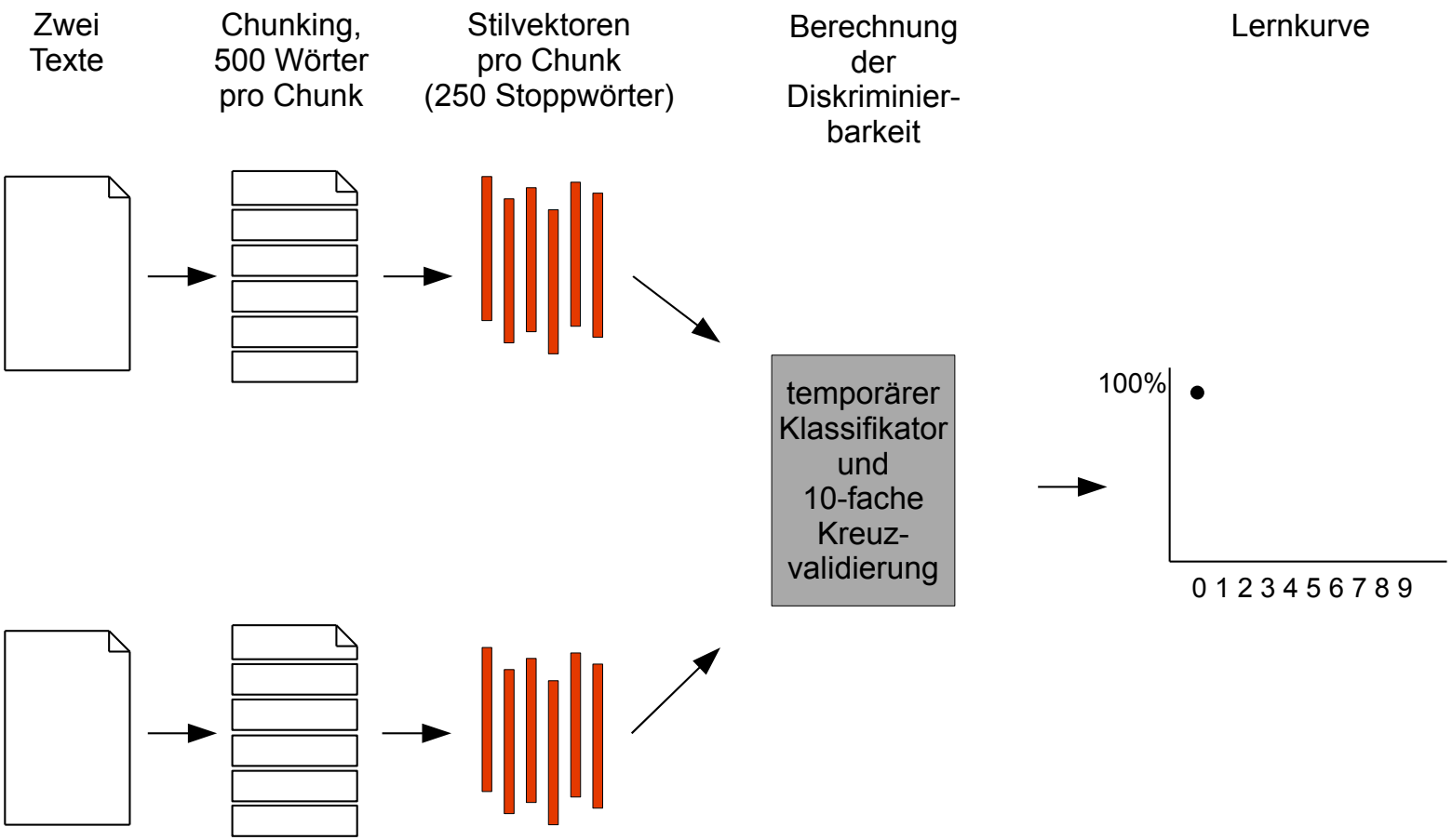
Stilvektoren pro Chunk (250 Stoppwörter)



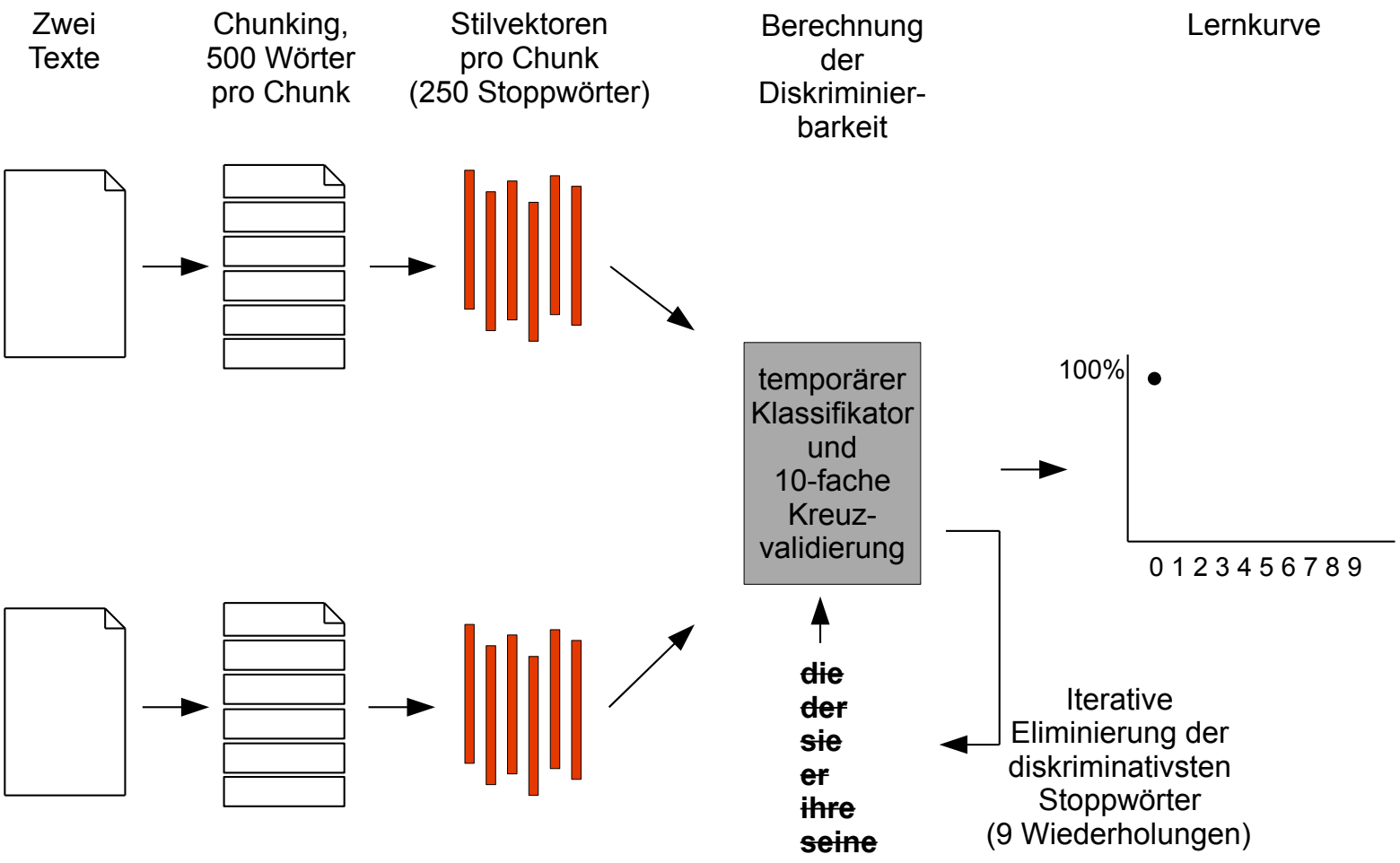
Unmasking


Zwei Texte
Chunking, 500 Wörter pro Chunk
Stilvektoren pro Chunk (250 Stoppwörter)
Berechnung der Diskriminier-barkeit
temporärer Klassifikator und 10-fache Kreuz-validierung

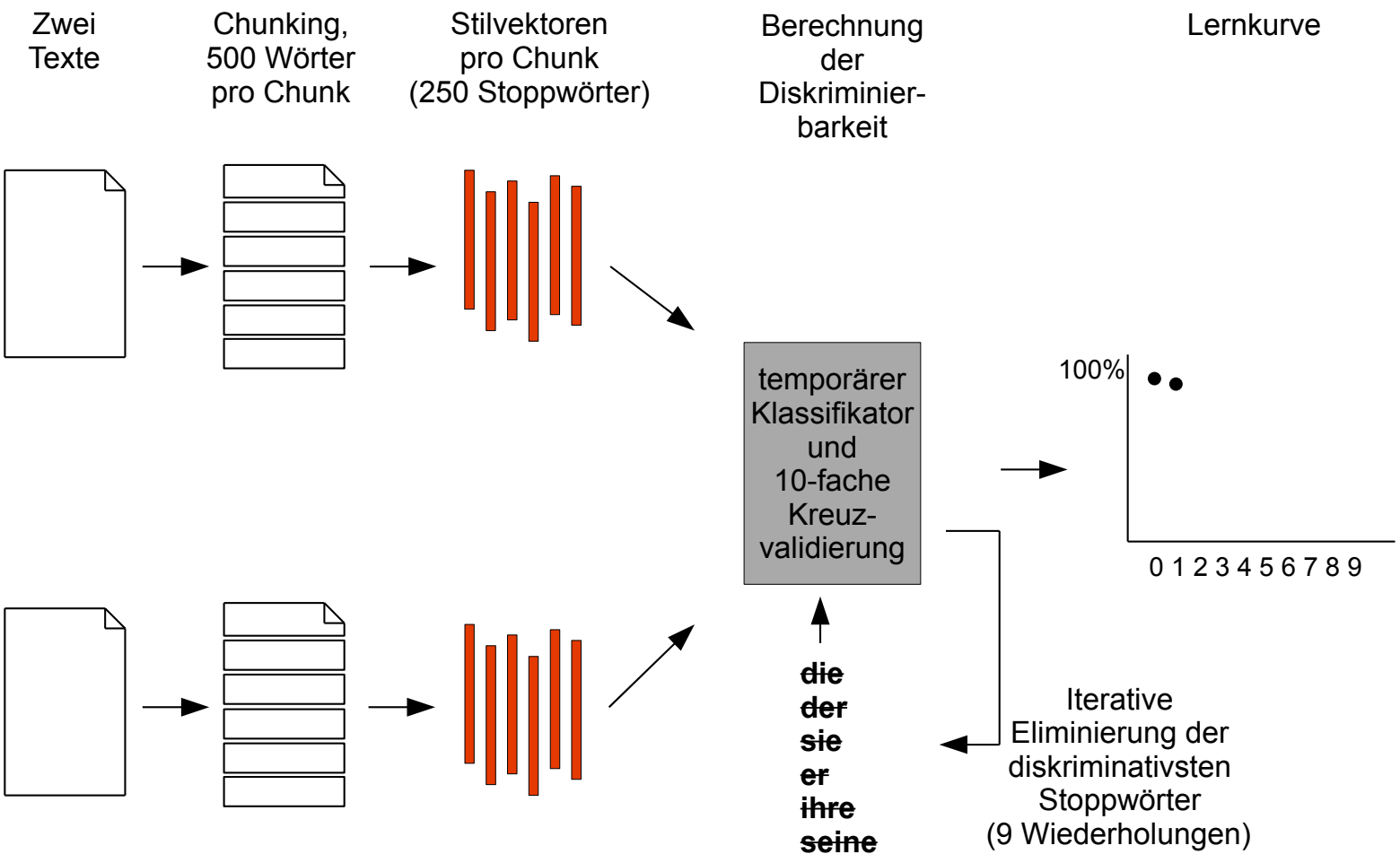
Unmasking


Zwei Texte
Chunking, 500 Wörter pro Chunk
Stilvektoren pro Chunk (250 Stoppwörter)
Berechnung der Diskriminier-barkeit
Lernkurve
temporärer Klassifikator und 10-fache Kreuz-validierung
100%
0 1 2 3 4 5 6 7 8 9

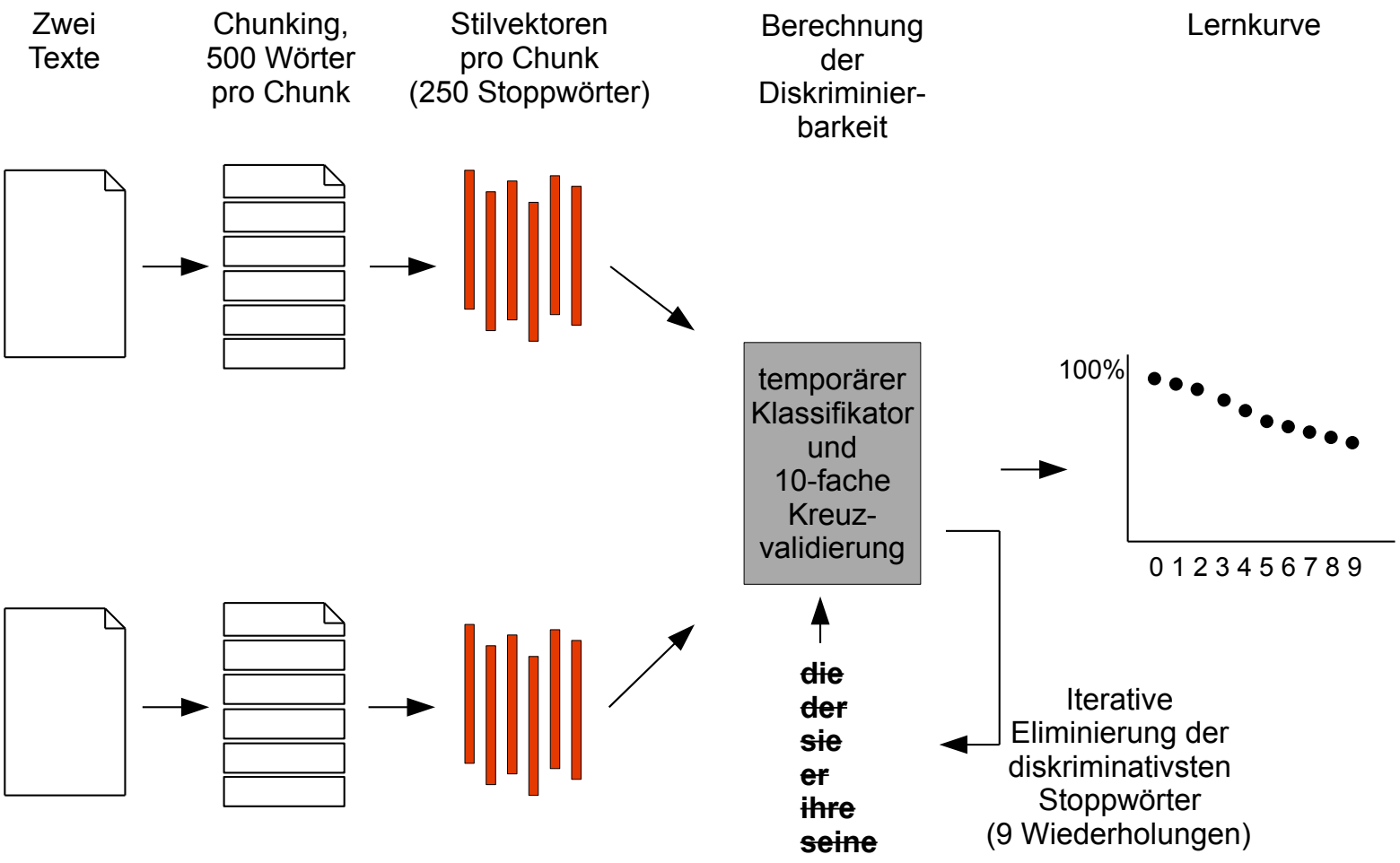
Unmasking


Zwei Texte
Chunking, 500 Wörter pro Chunk
Stilvektoren pro Chunk (250 Stoppwörter)
Berechnung der Diskriminier-barkeit
Lernkurve
temporärer Klassifikator und 10-fache Kreuz-validierung
100%
0 1 2 3 4 5 6 7 8 9
die
der
sie
er
ihre
seine
Iterative Eliminierung der diskriminativsten Stoppwörter (9 Wiederholungen)

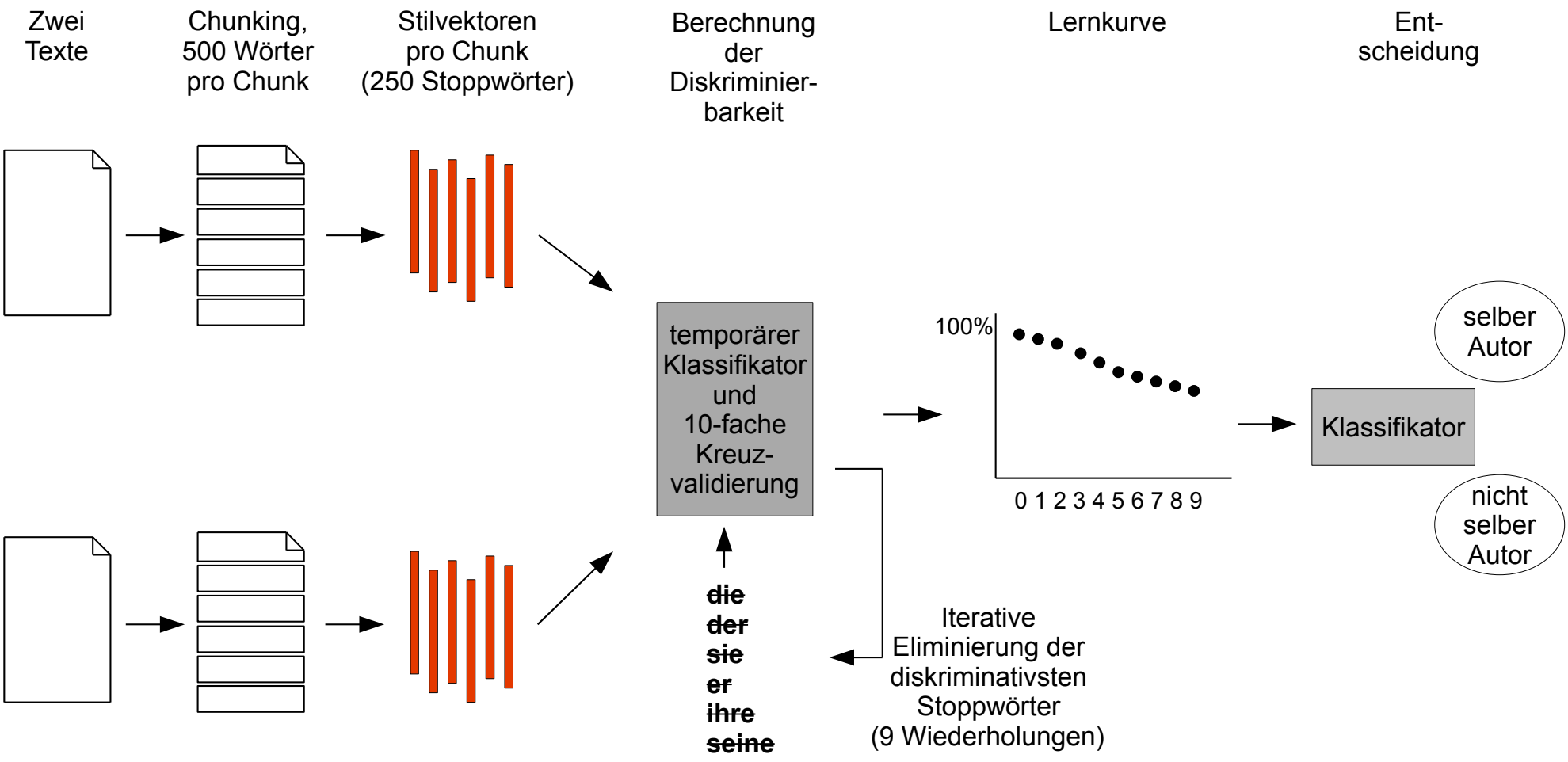
Unmasking


Zwei Texte
Chunking, 500 Wörter pro Chunk
Stilvektoren pro Chunk (250 Stoppwörter)
Berechnung der Diskriminier-barkeit
Lernkurve
temporärer Klassifikator und 10-fache Kreuz-validierung
100%
0 1 2 3 4 5 6 7 8 9
die
der
sie
er
ihre
seine
Iterative Eliminierung der diskriminativsten Stoppwörter (9 Wiederholungen)

Unmasking


Zwei Texte
Chunking, 500 Wörter pro Chunk
Stilvektoren pro Chunk (250 Stoppwörter)
Berechnung der Diskriminier-barkeit
Lernkurve
temporärer Klassifikator und 10-fache Kreuz-validierung
100%
0 1 2 3 4 5 6 7 8 9
die
der
sie
er
ihre
seine
Iterative Eliminierung der diskriminativsten Stoppwörter (9 Wiederholungen)

Unmasking

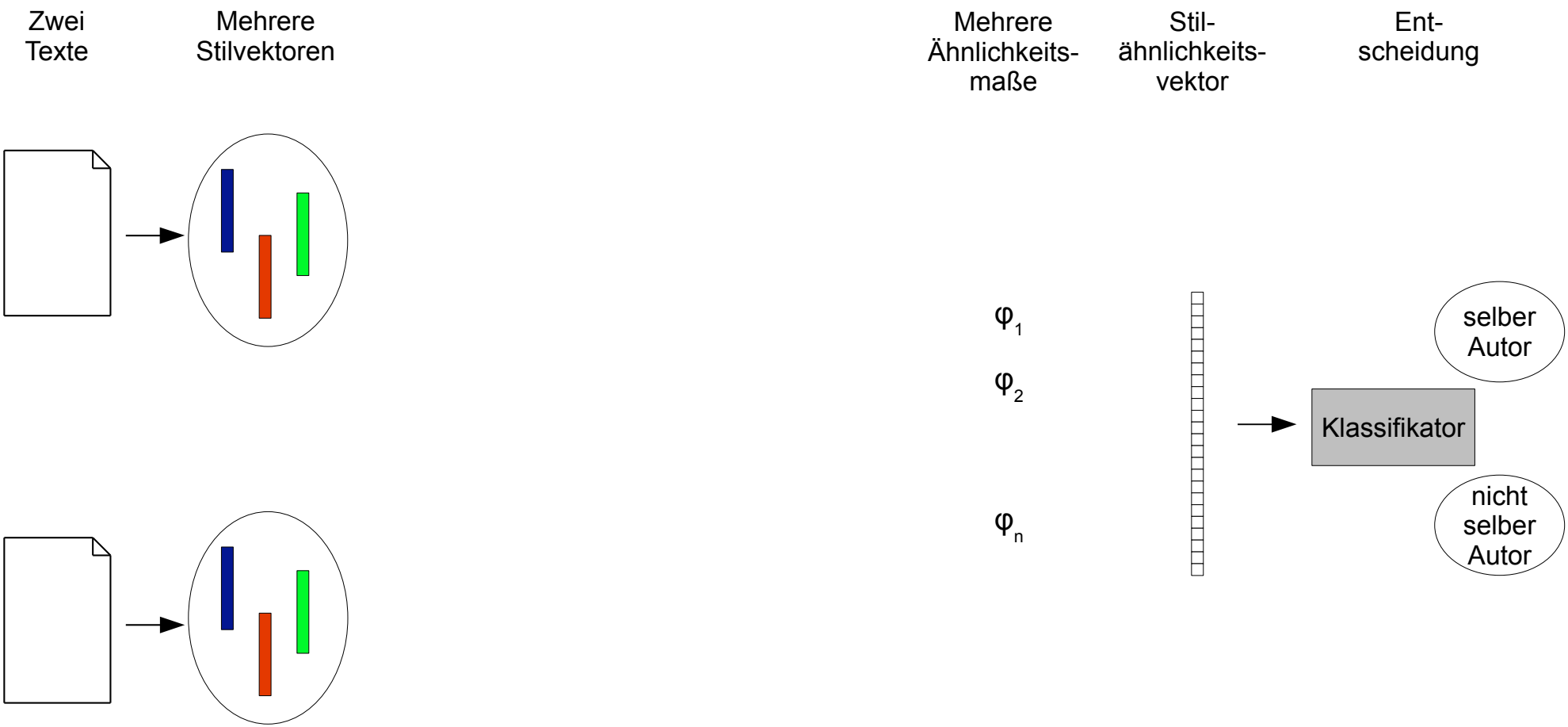

Zwei Texte
Chunking, 500 Wörter pro Chunk
Stilvektoren pro Chunk (250 Stoppwörter)
Berechnung der Diskriminier-barkeit
Lernkurve
Ent-scheidung
temporärer Klassifikator und 10-fache Kreuz-validierung
100%
0 1 2 3 4 5 6 7 8 9
Klassifikator
selber Autor
nicht selber Autor
die
der
sie
er
ihre
seine
Iterative Eliminierung der diskriminativsten Stoppwörter (9 Wiederholungen)

ST-Unmasking

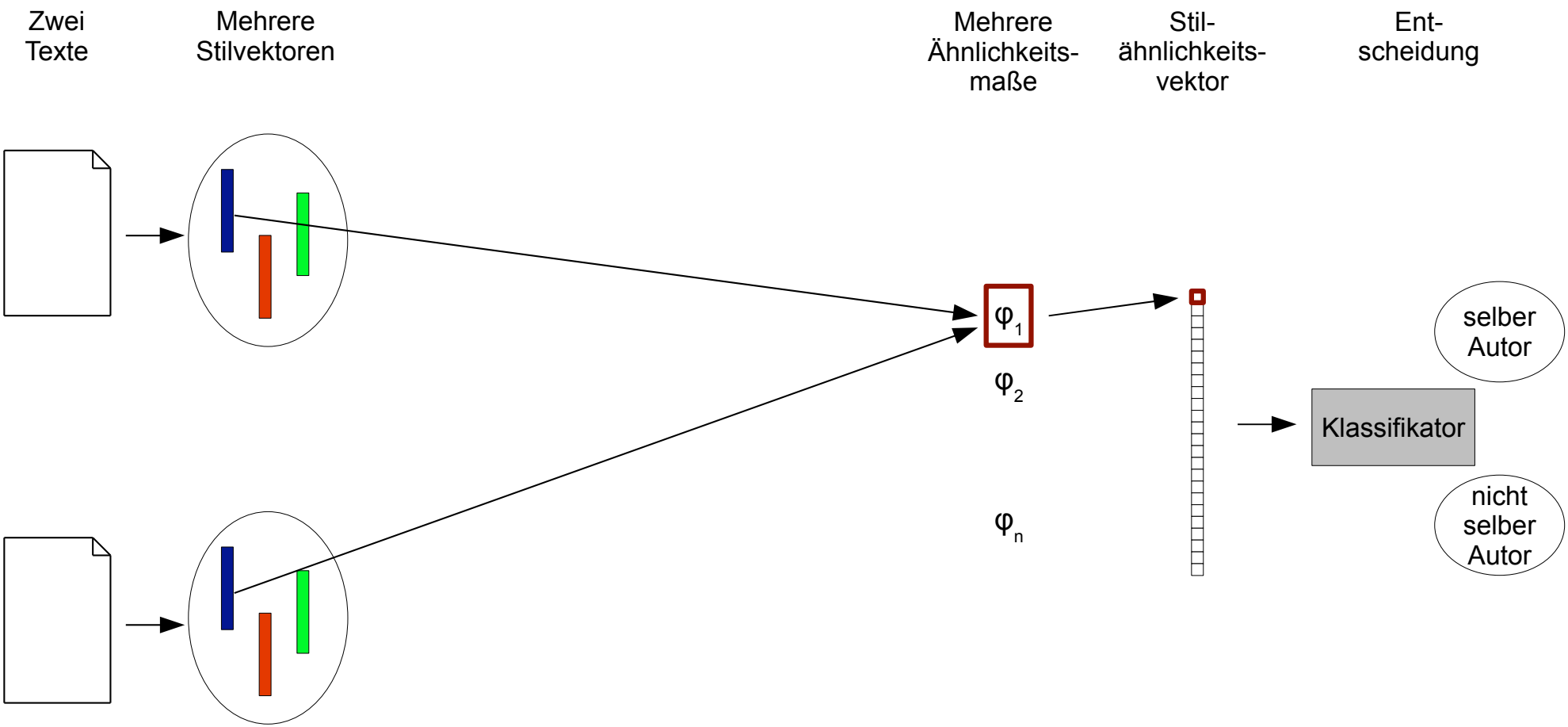
ST-Unmasking

ST-Unmasking

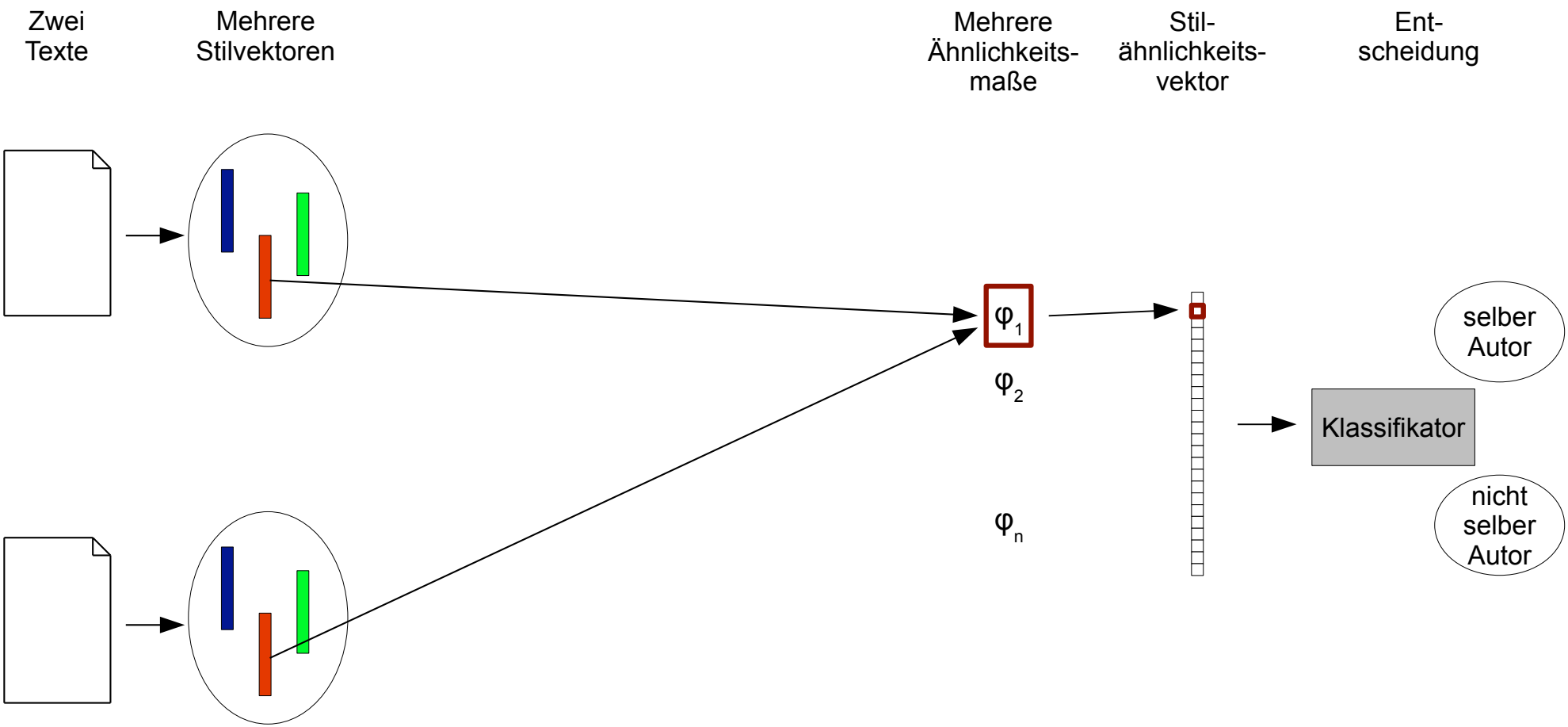
ST-Unmasking



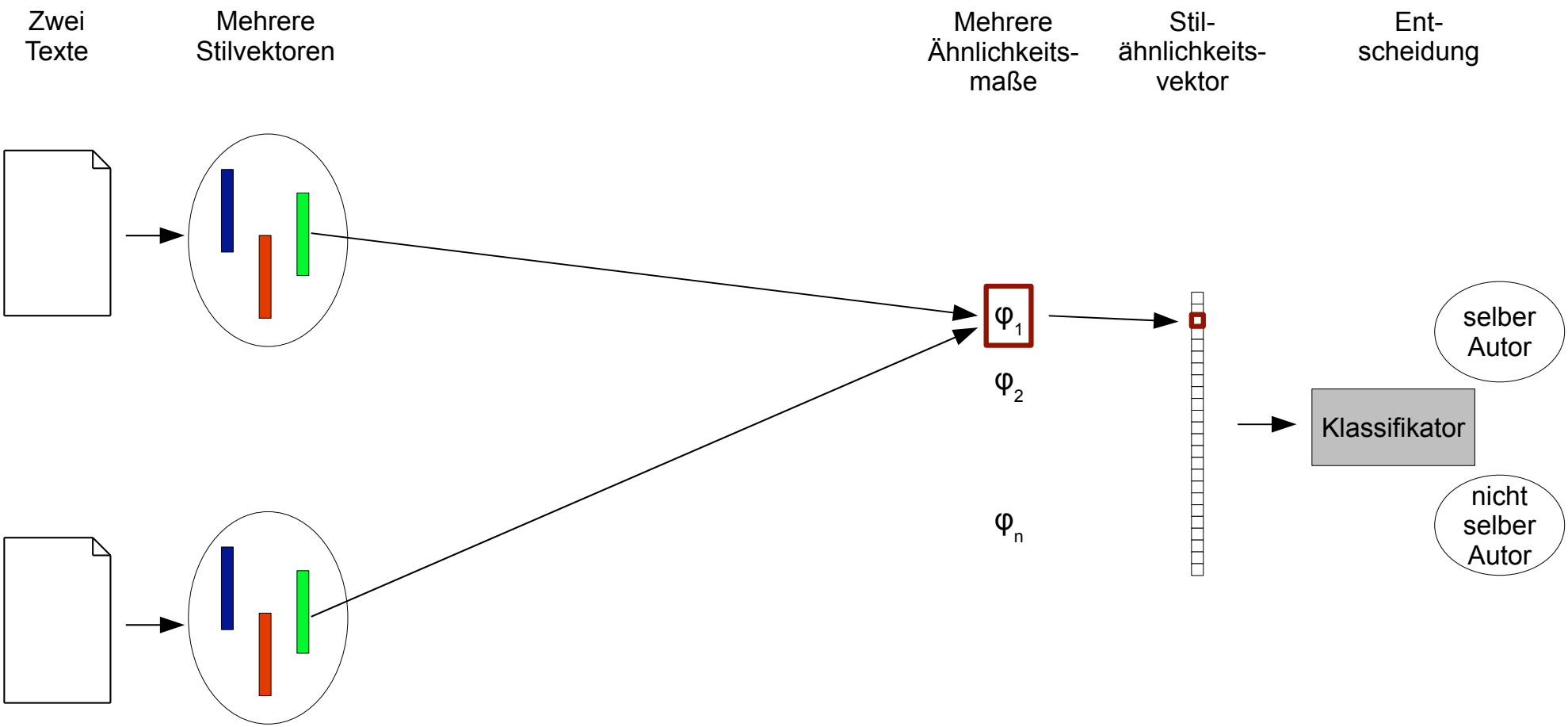
ST-Unmasking



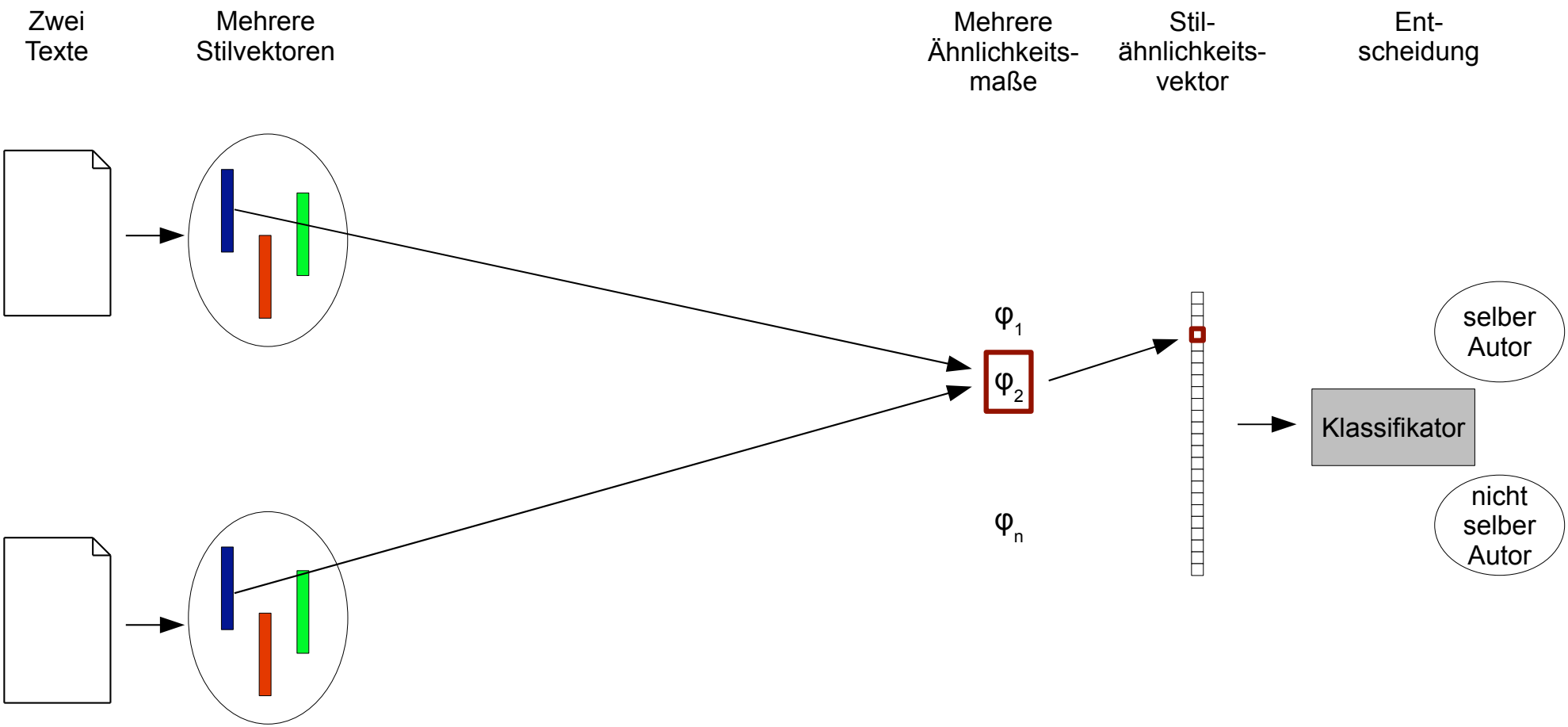
ST-Unmasking



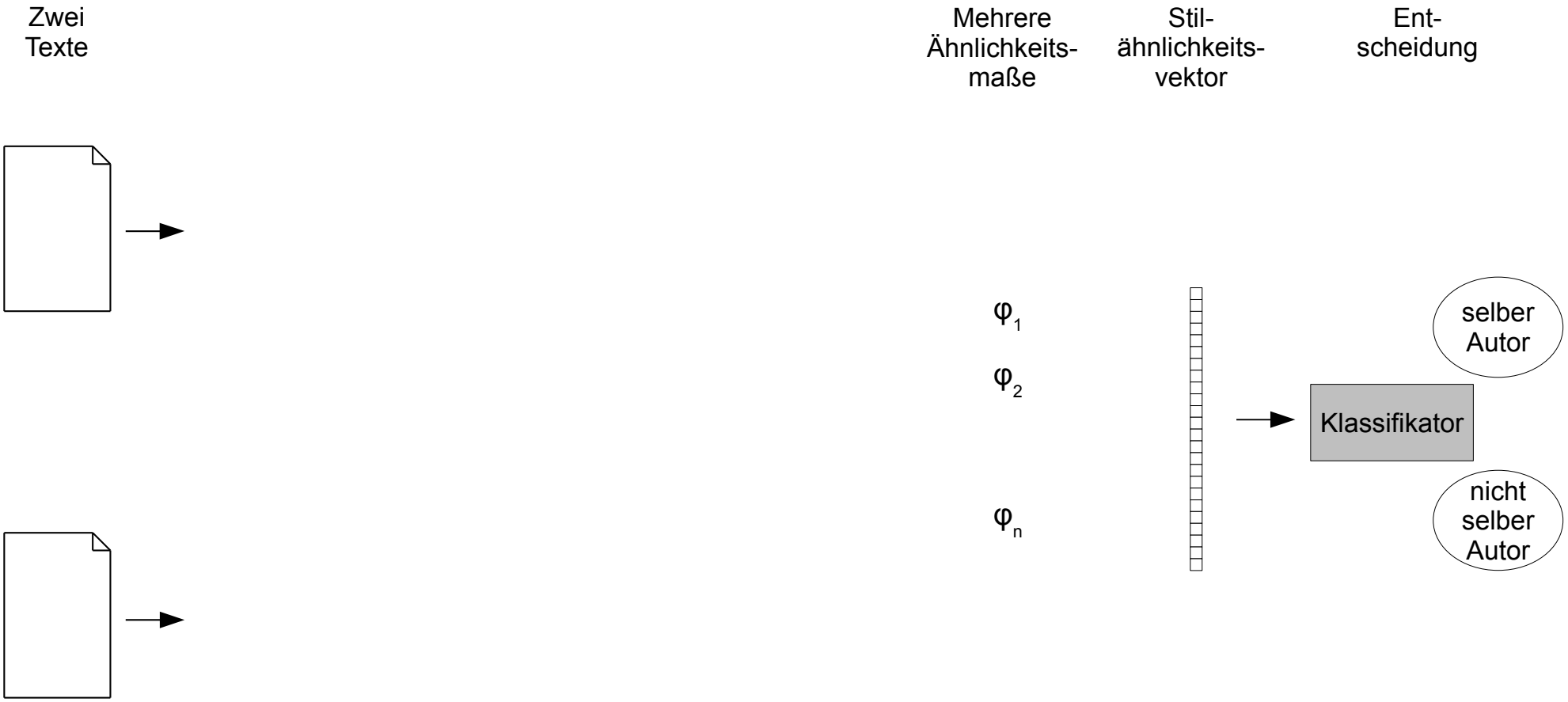
ST-Unmasking



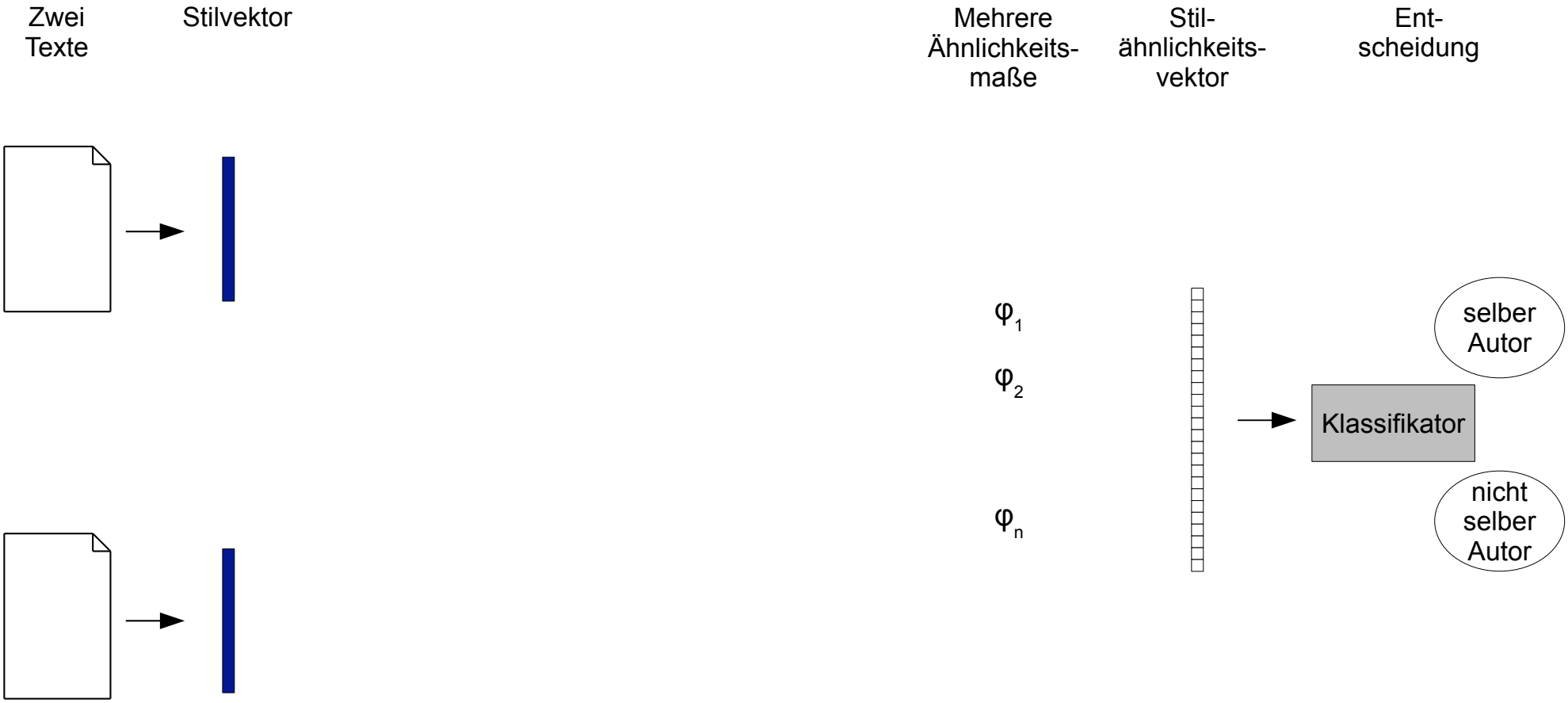
ST-Unmasking



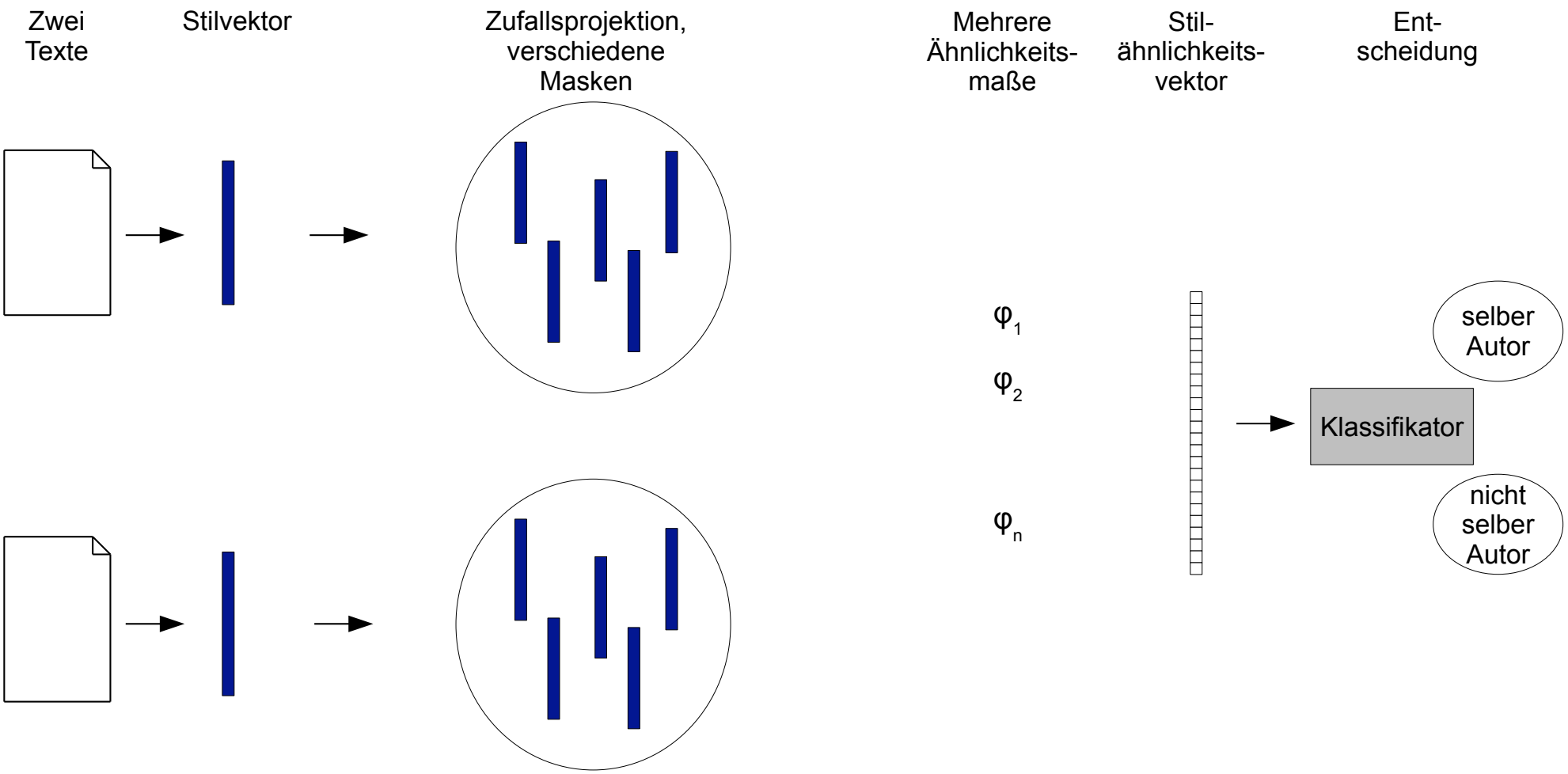
ST-Unmasking (2)



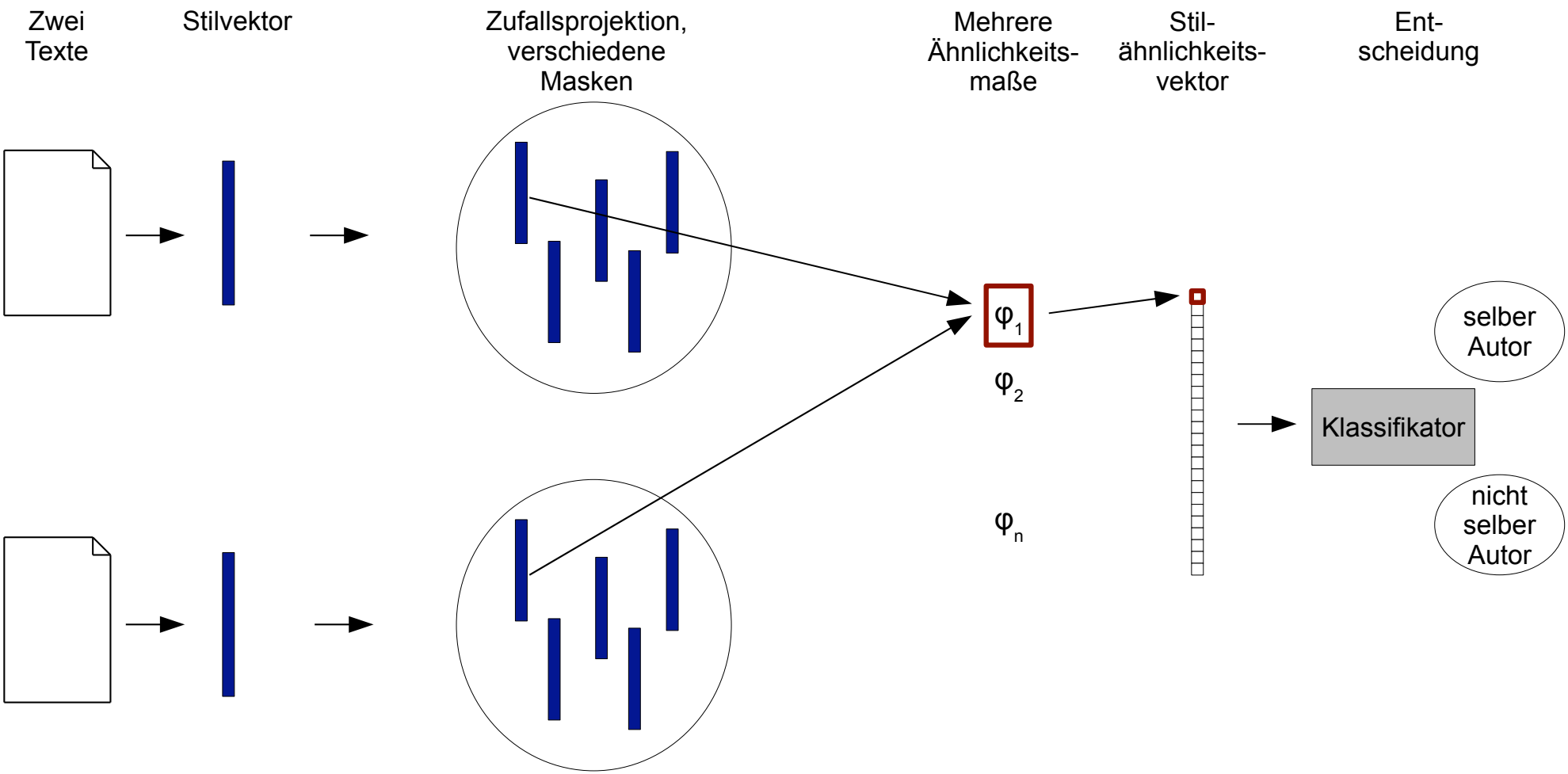
ST-Unmasking (2)



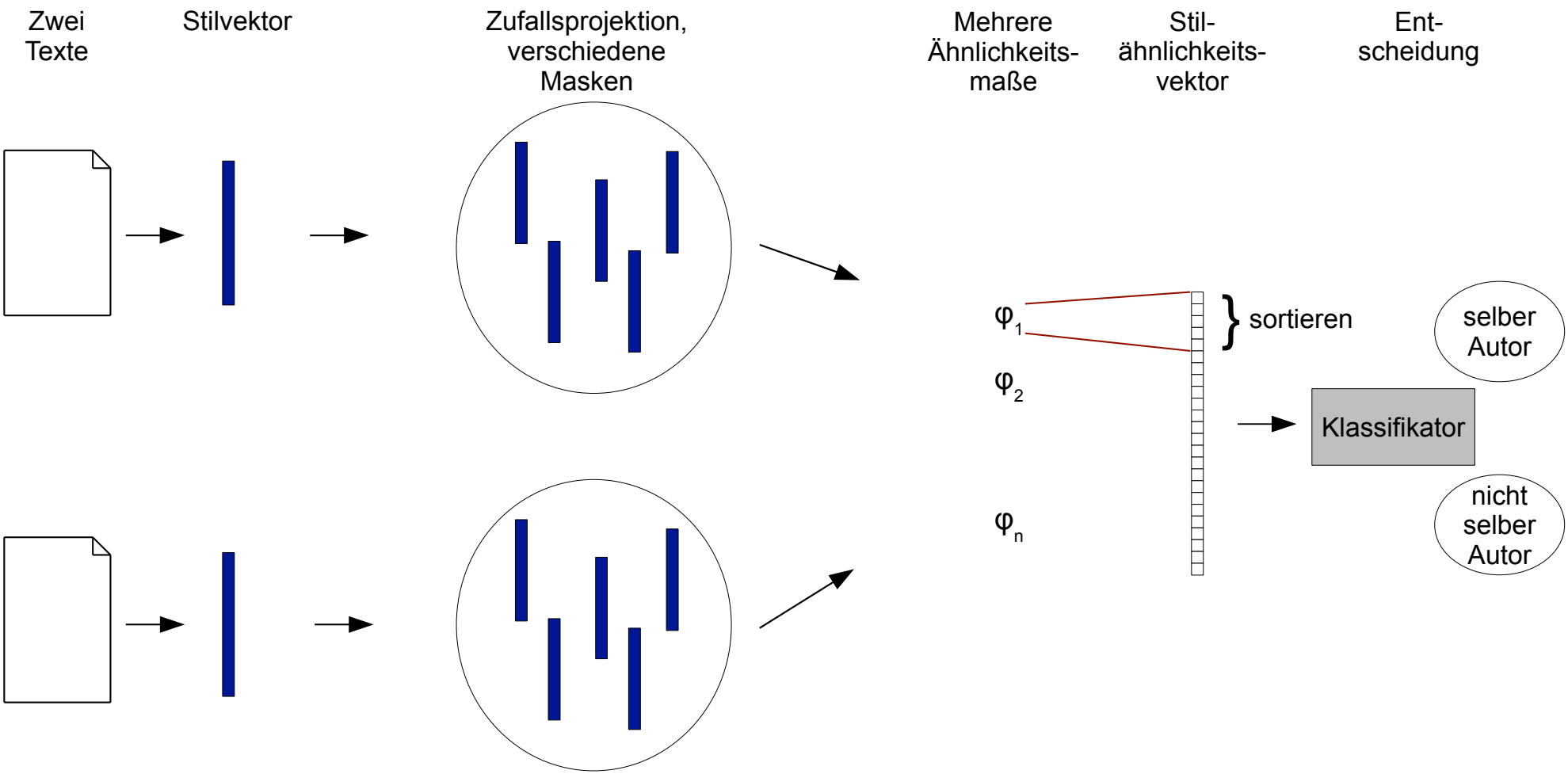
ST-Unmasking (2)



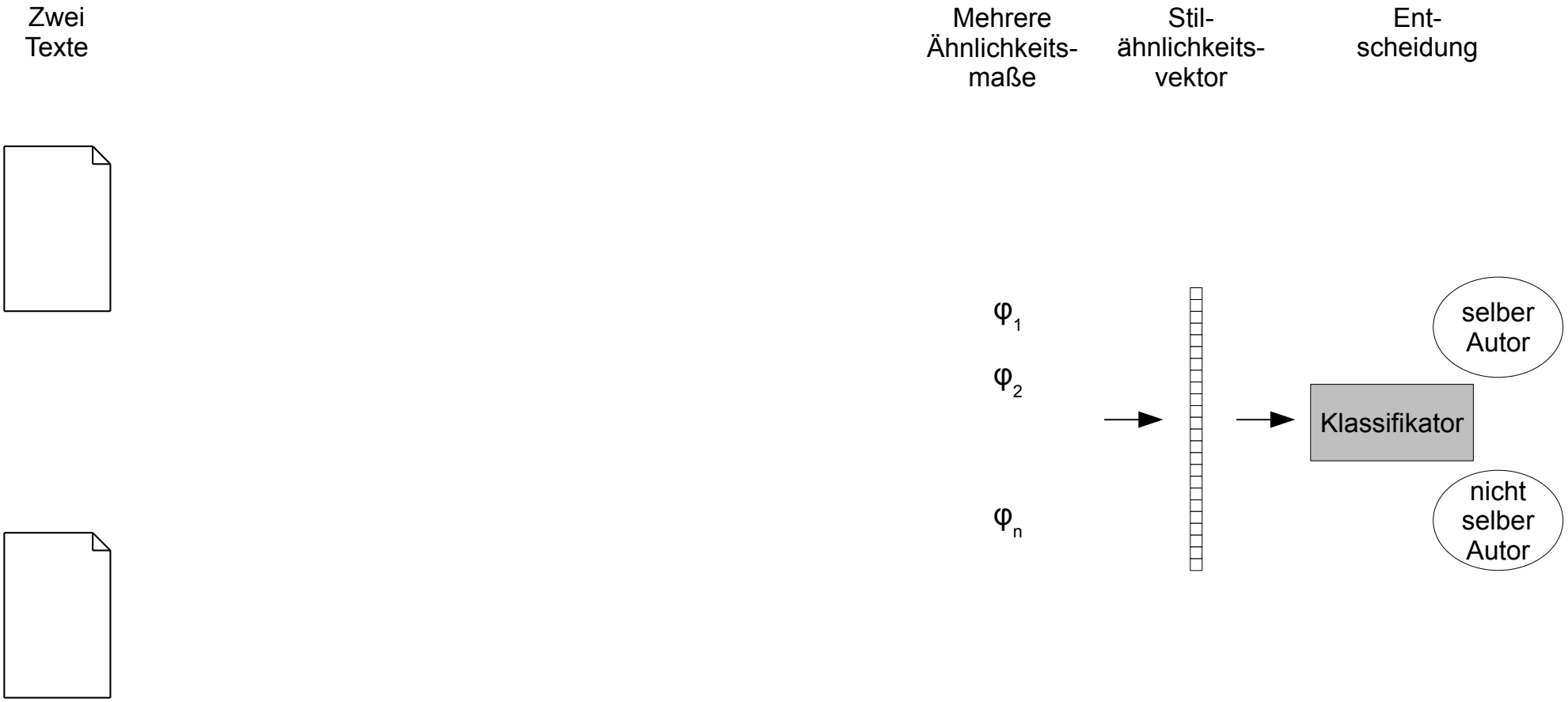
ST-Unmasking (2)



ST-Unmasking (2)

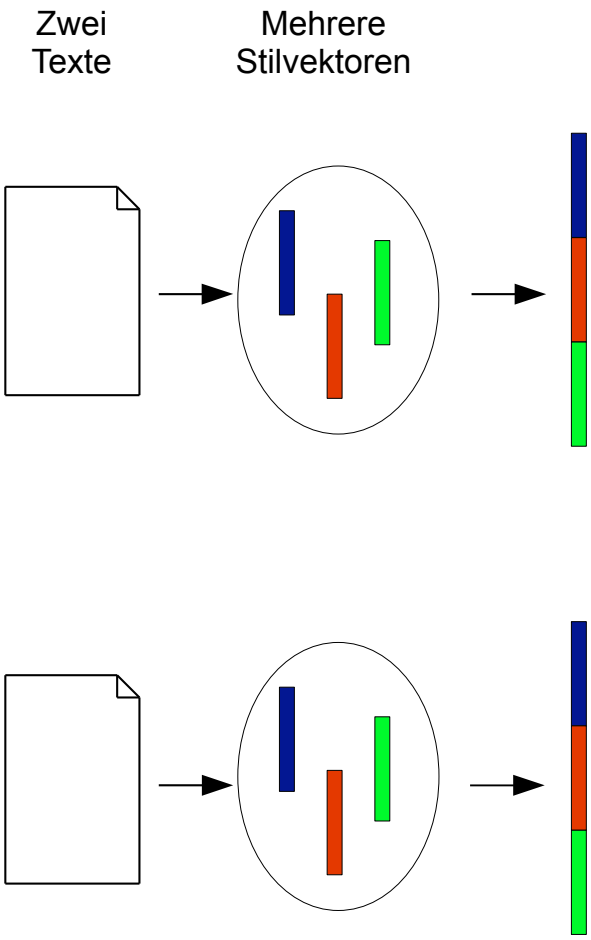


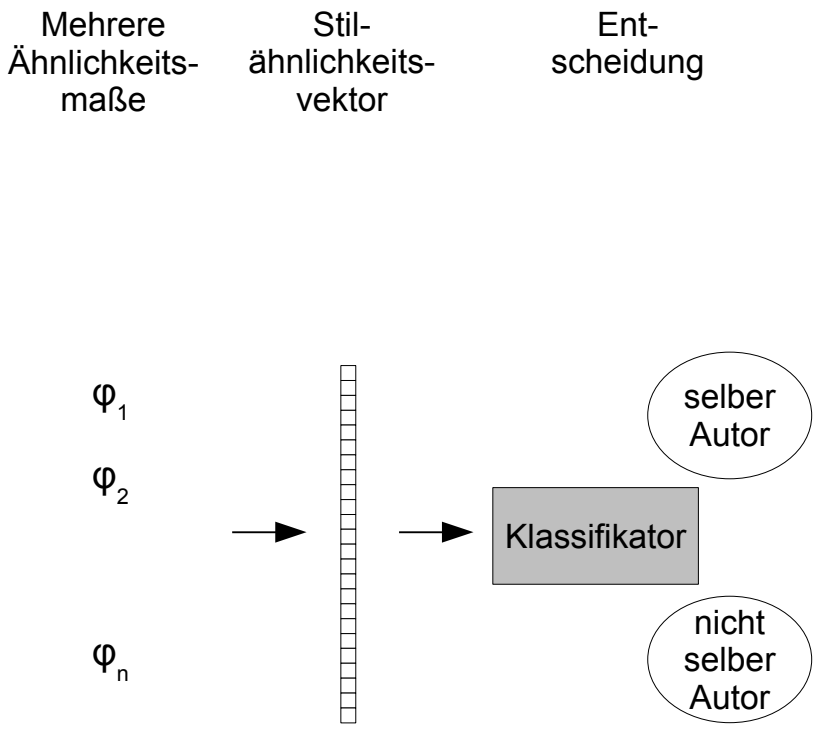
ST-Unmasking (3)



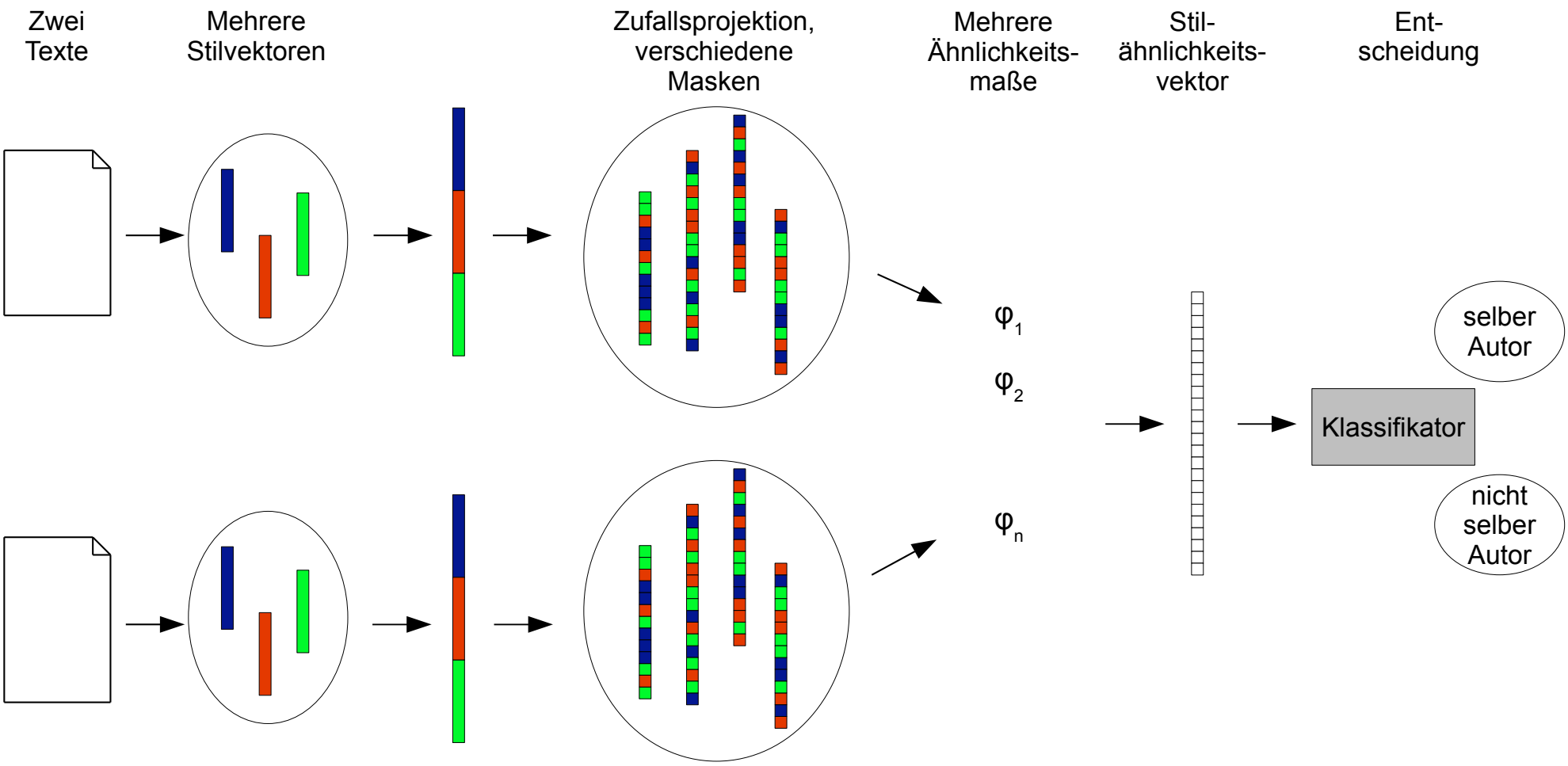
ST-Unmasking (3)

ST-Unmasking (3)





ST-Unmasking (3)

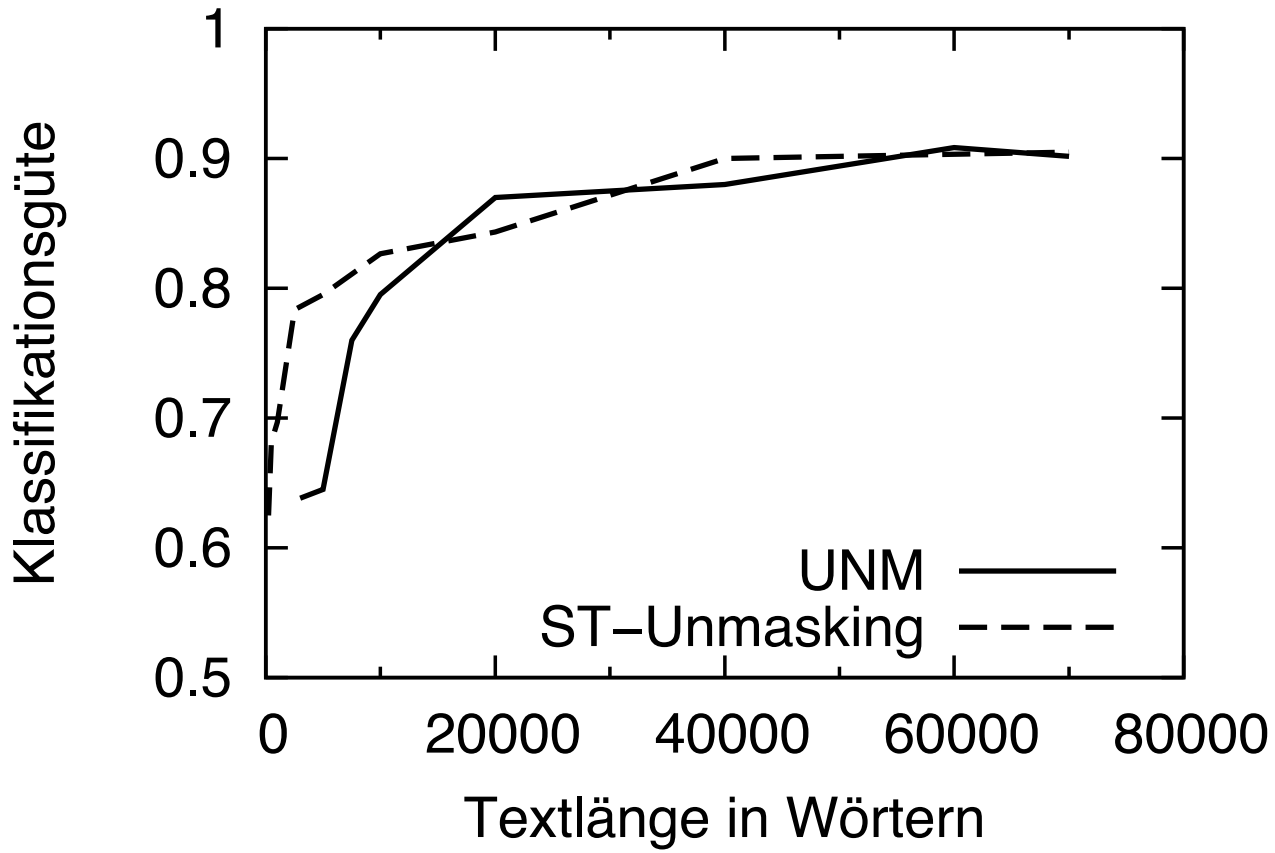


Evaluierung - Korpus

- 20.000 englische Bücher von 7.400 Autoren
- aus Projekt Gutenberg
- 200 Textpaare Trainingsmenge
- 600 Textpaare Testmenge

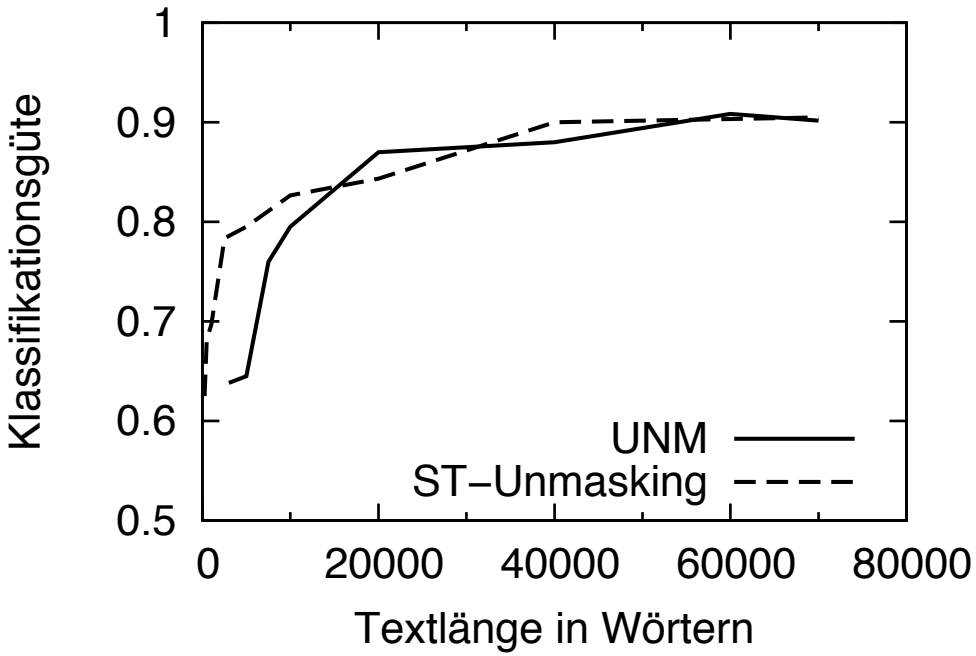
Evaluierung

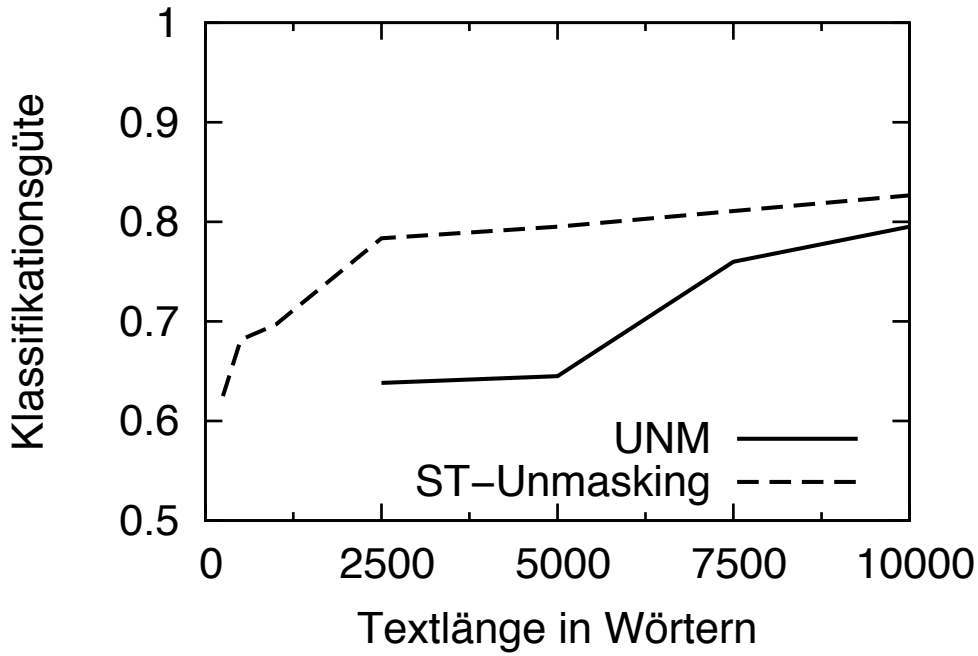
- Vergleich von Unmasking und ST-Unmasking
  - ST-Unmasking im Bereich kurzer Texte um etwa 15 Prozentpunkte besser
  - im Bereich längerer Texte vergleichbare Klassifikationsgüte



Evaluierung

- Vergleich von Unmasking und ST-Unmasking
  - ST-Unmasking im Bereich kurzer Texte um etwa 15 Prozentpunkte besser
  - im Bereich längerer Texte vergleichbare Klassifikationsgüte





Zusammenfassung

- Neu entwickelte Stilmerkmale
- Neuer Algorithmus zur paarweisen Autorenschaftsverifikation
- Evaluierung und Vergleich mit State-of-the-art Algorithmus
- Steigerung der Klassifikationsgüte um 15 Prozentpunkte bei 2500 Wörtern
- Vergleichbare Güte bei langen Texten

Literatur

1. **Koppel, M. und J. Schler:** *Authorship verification as a one-class classification problem.* In: Proceedings of the twenty-first international conference on Machine learning, Seite 62. ACM, 2004.