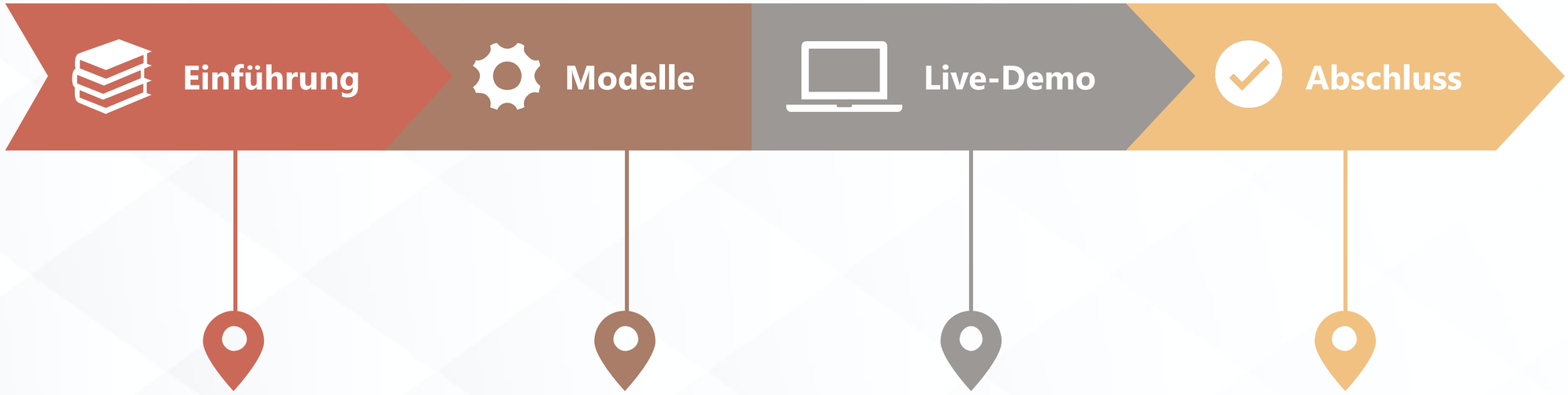
Integrationsseminar Autorenprofiling mit Machine Learning Methoden

Anabel Lilja, Bjarne Gerdes, Johannes Deufel und Simone Marx

Agenda



Einführung
- Relevanz
- Related Work

Modelle
- Vorgehen
- Modelle
- Ergebnisse

Live-Demo
- API
- Frontend

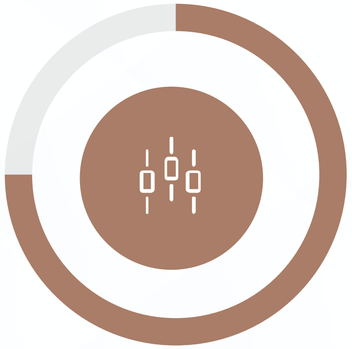
Abschluss
- Fazit
- Ausblick

Einführung

Stilometrie

Untersuchung von Sprachstilen mithilfe statistischer Mittel

Jeder Autor besitzt einen (unbewussten) individuellen Schreibstil



Schreibstile lassen sich durch die konsistente Verwendung bestimmter Muster statistisch abgrenzen

Relevanz

Forensik

Überführung von
verdächtigen Personen

Marketing

Kundenbindung
Kundengewinnung
Personalisiertes Marketing

Bibliotheken

Zuordnung von
Autoren zu Werken
Historische Rückschlüsse
Über Epochen/
Regionen/Gesellschaften

Akademik

Plagiat Erkennung

Related Work

Devi et al.
Datensatz: SMS-AP-18
Algorithmus: SVM

Geschlecht:
85,7%
Accuracy

Alter:
64,3%
Accuracy

Modaresi et al.
Datensatz: Twitter, PAN2014
Algorithmus: Logistic Regression

Geschlecht:
75,6%
Accuracy

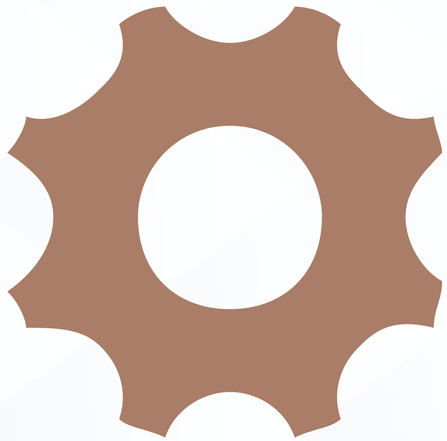
Alter:
51,7%
Accuracy

Azmi & Al-Ghadir
Datensatz: Social Media Posts Saudi-Arabien
Algorithmus: SVM, KNN

Geschlecht SVM:
87,3%
Accuracy

Geschlecht KNN:
93,1%
Accuracy

Modelle



Data Cleaning

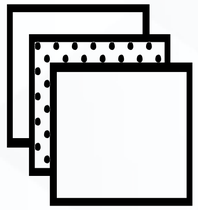
Feature Engineering

- Anzahl der Zeichen, Emails, Zahlen, Groß-/Kleinbuchstaben, URLs und Sonderzeichen im Text
- Anzahl der verschiedenen verwendeten Worte im Text
- Durchschnitt der Buchstaben pro Wort und pro Satz
- Durchschnitt der Worte pro Satz

- Varianz der Wortlänge, Varianz der Buchstaben und Worte pro Satz
- Anteil der Großbuchstaben in dem Satz mit dem höchsten Anteil an Großbuchstaben
- Lemmatisierung

Arten von Features

Vorhandene Arten von Features nach dem Feature Engineering

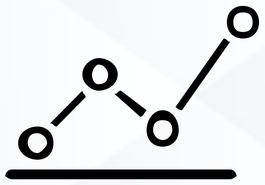


Kategoriale Variable

K

B

Vorverarbeiteter Text

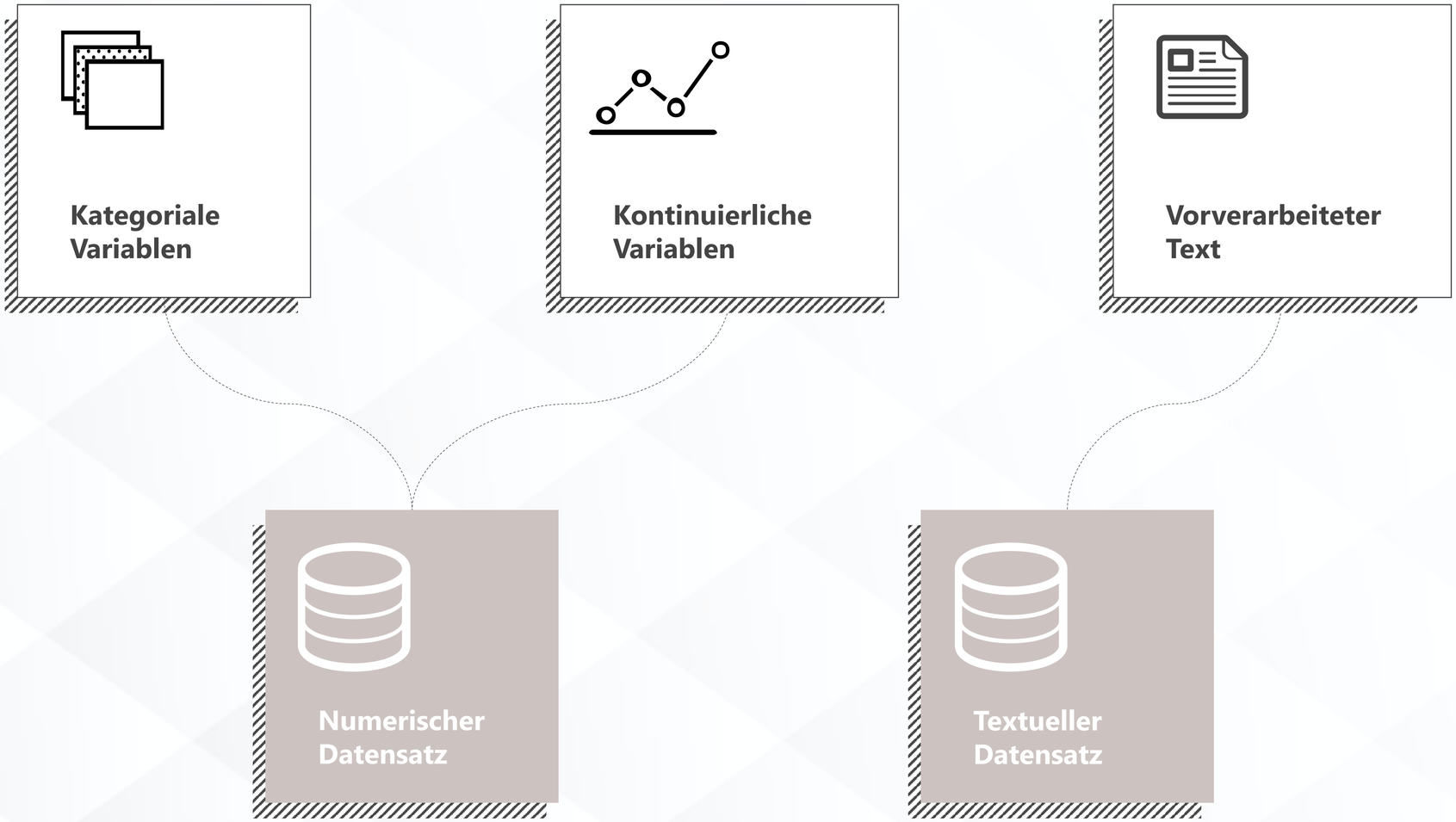


Kontinuierliche Features

C

Aufteilen der Features

Trennung des Textes von verbleibenden numerischen Features



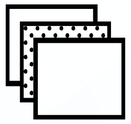
Numerische Transformation

Verarbeitung der nicht-text Features

Gender
female
male
female
male

Gender = female	Gender = male
1	0
0	1
1	0
0	1

$$Z(x) = \frac{x - \mu}{\sigma}$$



One-Hot Encoding

Für jede Ausprägung einer kategorialen Variable im Datensatz wird eine Variable erstellt. Diese trägt entweder die Ausprägung 1 oder 0, wobei 1 indiziert, ob die Ausprägung der initialen Variable der Ausprägung der neuen Variable entspricht.

Z-Transformation

Alle kontinuierlichen Variablen sind mittels Z-Transformation standardisiert worden, so dass ihr Mittelwert 0 und die Standardabweichung 1 entspricht. Dies vermeidet numerische Instabilitäten.

Textuelle Transformation

Transformation der Texte in TF*IDF-Vektoren

TF*IDF

Term Frequency

Relativer Anteil eines Terms in einem Dokument

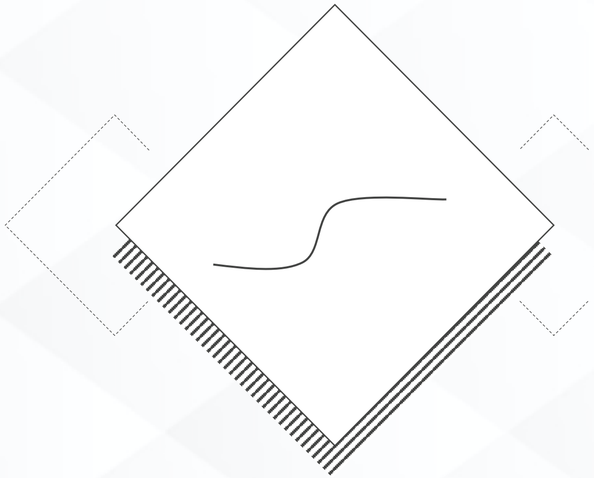
*

Inverse Document Frequency

Inverser Relativer Anteil an Dokumenten, die diesen Term enthalten

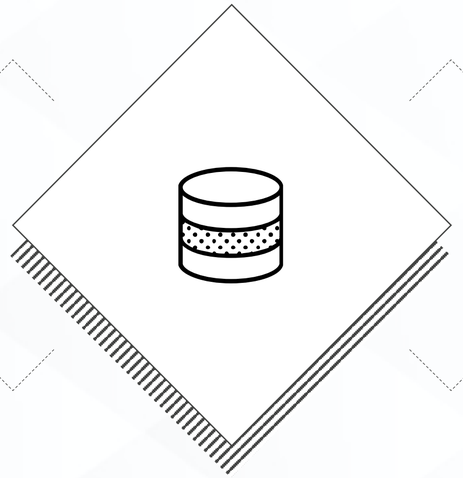
Supervised Learning

Modelle die zur Zielwertvorhersage verwendet worden sind



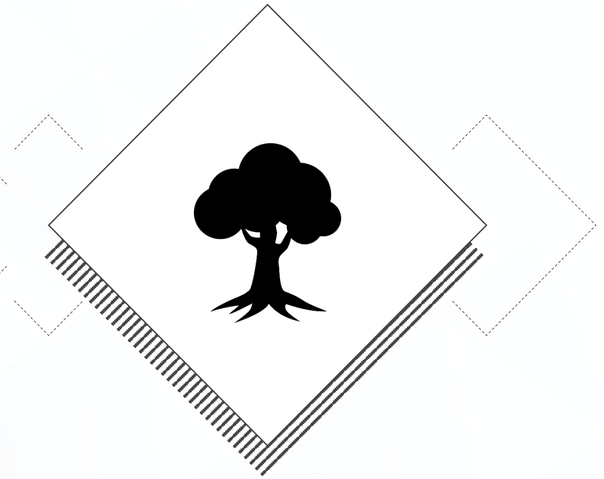
Logistische Regression

Die Logistische Regression kann mittels One-Versus-Rest-Verfahren für univariate und multinominale Klassifikation verwendet werden.



Support Vector Machine

Die Support Vector Machine bestimmt Entscheidungsgrenzen, so dass die Differenz zwischen den Klassen maximiert wird.

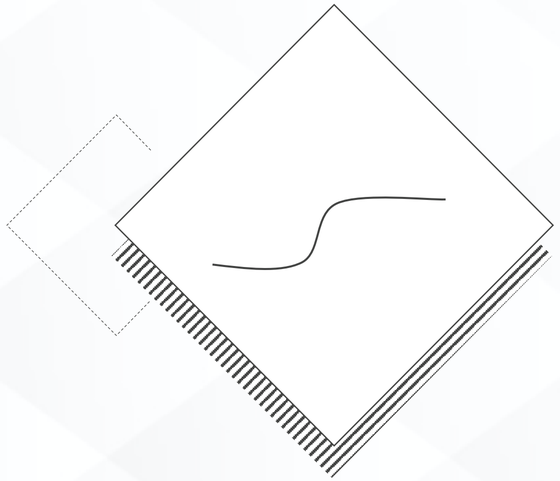


XGBoost

Baumbasierter Algorithmus, der auf Gradientenverfahren und Boosting zurückgreift.

Parametertuning

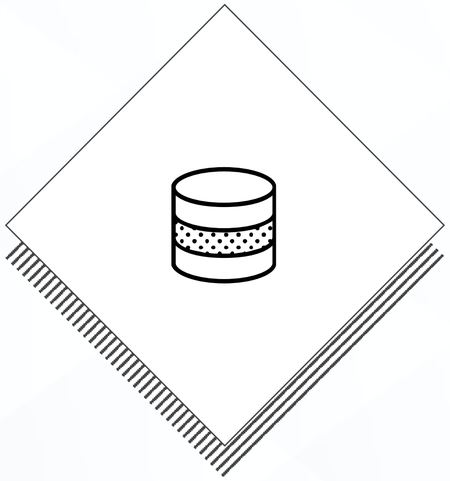
durch Cross Validation



Logistische Regression

Alpha: 0,001; 0,00001

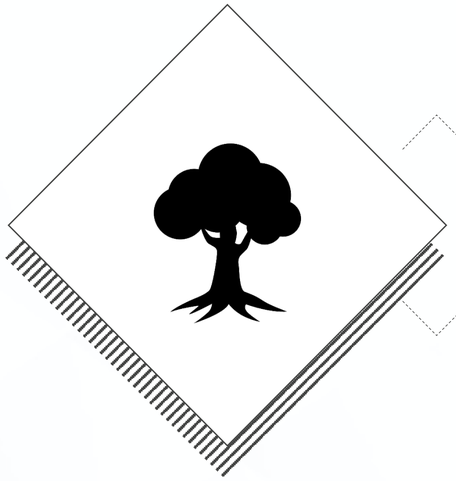
Penalty: elasticnet;
L1-Norm;
L2-Norm



Support Vector Machine

Loss: epsilon insensitive;
squared epsilon insensitive;
hinge;
squared hinge

C: 0,8; 1; 1,2



XGBoost

Learning Rate: 0,1; 1; 1,5

Max Depth: 3; 6

N Estimators: 400; 800

Modell Evaluation

Alter

Text: Linear Support Vector Regression, 49,109%

Numerisch: XGBoost, 50,891%

Negativer MSE: -4,648

Geschlecht

Text: Logistische Regression, 45,734%

Numerisch: XGBoost, 54,266%

F1-Score: 80,30%

Thema

Text: Linear Support Vector Classification, 38,950%

Numerisch: XGBoost, 61,050%

F1-Score: 39,307%

Sternzeichen

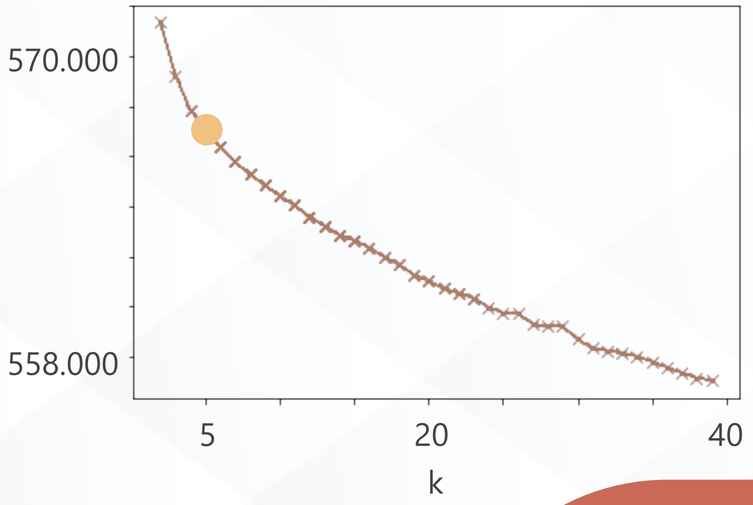
Text: logistische Regression, 46,390%

Numerisch: XGBoost, 53,710%

F1-Score: 42,20%

Unsupervised Learning

k-means



Textuell

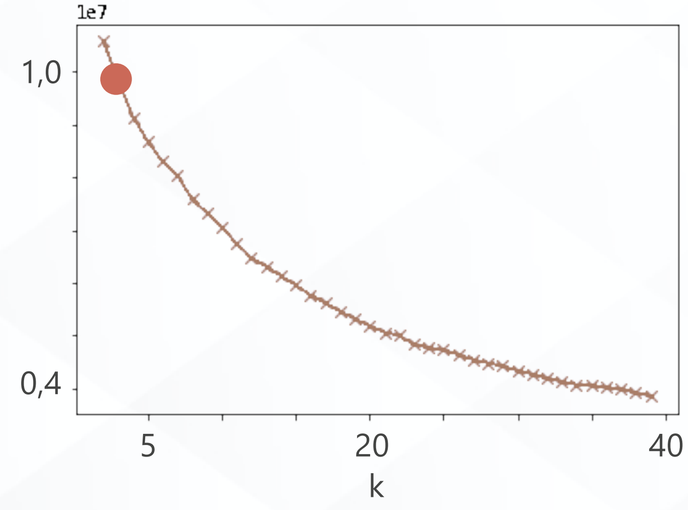
Basierend auf Wortwahl

“Negation-Lover”

Numerisch

Basierend auf stilistischen Eigenheiten

“Hobby Publisher”



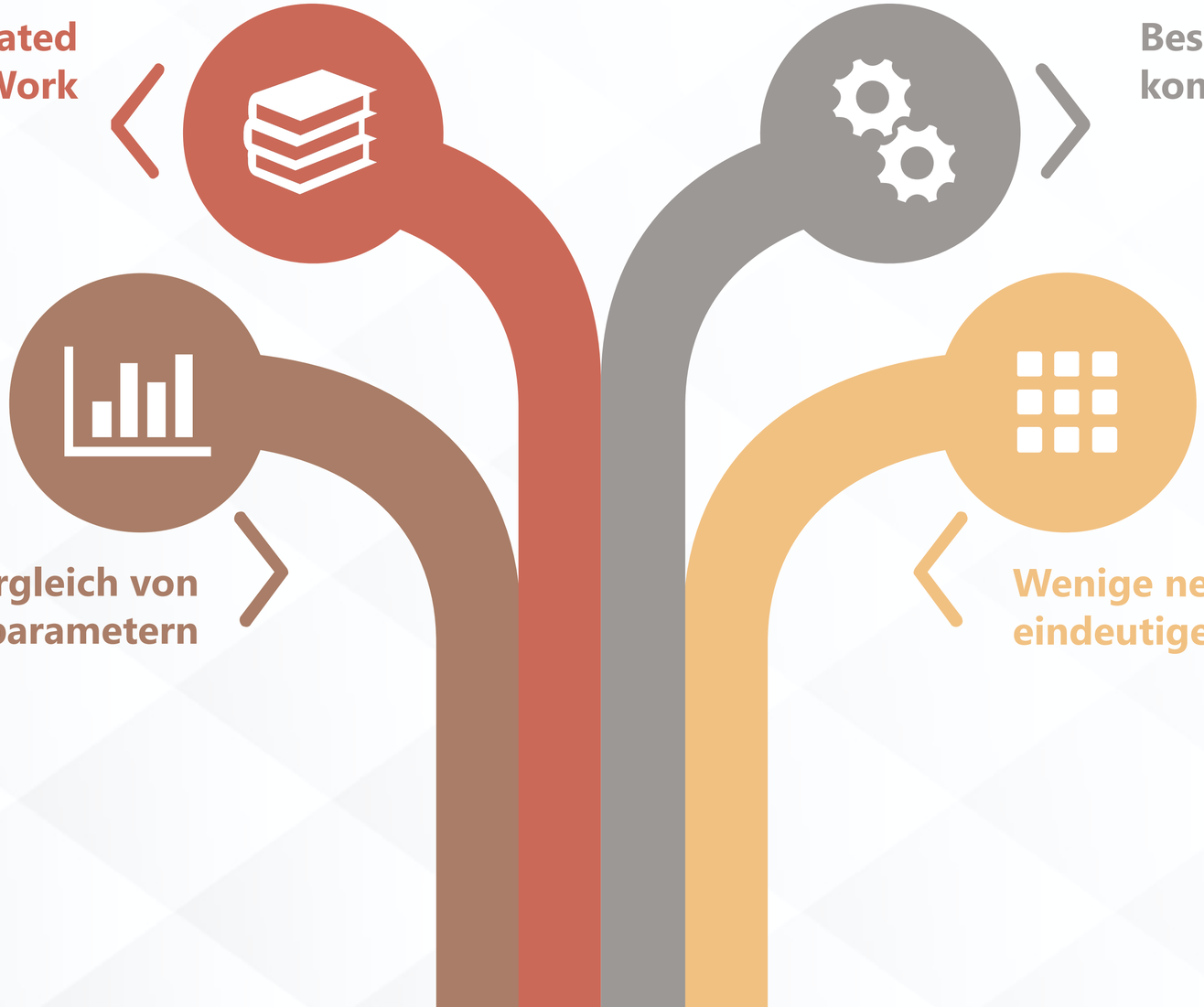
Live-Demo

Abschluss

Fazit

Modelle ähnlich gut wie Related Work

Bessere Ergebnisse durch kombinierte Modelle



Mehrwert durch breiten Vergleich von Algorithmen und Hyperparametern

Wenige neue Cluster, dafür eindeutige Merkmale

Ausblick

Diverserer und hochwertigerer Datensatz

Weitere Clusteringalgorithmen

Verwendung von Word-Embeddings