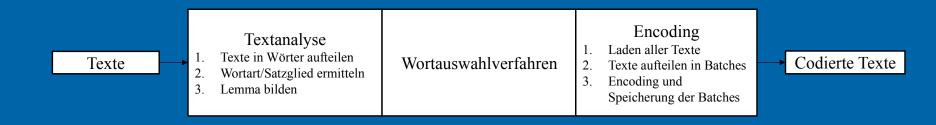


# Verteidigung der Bachelorarbeit

## Gegenüberstellung von Preprocessing Methoden der Textklassifizierung von Zeitungsmeldungen durch künstliche neuronale Netze



Von Erik Zimmermann 4.8.2021



# **Beschreibung des Titels**

- Textklassifizierung
  - Einsortieren von Texten in zuvor definierte Sparten [1]
  - z.B.: Aufteilung von Zeitungen in Rubriken
- Preprocessing
  - Umwandlung Daten in validen Netzinput [2]
  - z.B.: Textsäuberung, Umwandlung in numerische Darstellung

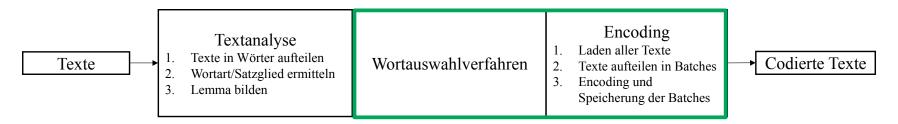


# Aufgabenstellung

- Preprocessing in verschiedene Schritte unterteilen
- Schritte verändern zur Kombinationsbildung
- Texte nach Kombinationen codieren
- Durch Convolutional Neural Network klassifiziert
- Gegenüberstellung der Ergebnisse
- Artikel verschiedener Zeitungen
- In neun Sparten unterteilt



# Aufbau des Preprocessings



- Textanalyse:
  - Umwandlung von Text in Wortliste
  - Ermittlung von Worteigenschaften
- Wortauswahlverfahren: Eliminierung von Worten nach Vorgaben
- Encoding: Umwandlung in Netzinput



## Wortauswahlverfahren

- Keine Wortauswahl
- Auswahl nach Worttyp
  - Substantive, Verben
  - Schlüsselwörter
- Auswahl nach Satzglied
  - Subjekte, Prädikate
  - Essentiell für Satzbildung

- Auswahl nach TF-IDF
  - Wichtigkeit des Begriffes in Text
  - Begriffe mit TF-IDF über Grenzwert



# **Encoding**

- Normalized Bag of Words
  - Darstellung Anteil eines Wortes am Text
- Wortrepräsentation durch Vektor
  - Ersetzung Wort mit Vektor
  - Vorkommen von Wort pro Sparte
- Ordinal Encoding
  - Ersetzung Wort mit Index
  - Indexbestimmung durch Wörterbuch

# The human builds the house for shelter. The human lives in the house.

| the     | 0.3077 |
|---------|--------|
| human   | 0.1538 |
| build   | 0.0769 |
| house   | 0.1538 |
| for     | 0.0769 |
| shelter | 0.0769 |
| live    | 0.0769 |
| in      | 0.0769 |

| 3 | 3                     | 2                               |  | the  |
|---|-----------------------|---------------------------------|--|--|
| 2 | 0                     | 0                               |  | human  |
| 2 | 1                     | 0                               |  | build  |
| 1 | 5                     | 1                               |  | house  |
| 4 | 2                     | 4                               |  | for  |
| 1 | 0                     | 0                               |  | shelter  |
| 1 | 1                     | 1                               |  | live   |
| 2 | 3                     | 2                               |  | in   |
|   | 2<br>2<br>1<br>4<br>1 | 2 0<br>2 1<br>1 5<br>4 2<br>1 0 | 2 0 0<br>2 1 0<br>1 5 1<br>4 2 4<br>1 0 0<br>1 1 1 | 2 0 0<br>2 1 0<br>1 5 1<br>4 2 4<br>1 0 0<br>1 1 1 |

| 3 | 2 | 2 | 3 | 1 | 4 | 1 |
|---|---|---|---|---|---|---|
| 3 | 0 | 1 | 3 | 5 | 2 | ( |
| 2 | 0 | 0 | 2 | 1 | 4 | ( |
| 3 | 2 | 1 | 2 | 3 | 1 |   |
| 3 | 0 | 1 | 3 | 3 | 5 |   |
| 2 | 0 | 1 | 2 | 2 | 1 |   |

| 0 | 1 | 2 | 0 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| 0 | 1 | 6 | 7 | 0 | 3 |   |



# Ablauf des Experimentes

- Trainingsset (9000 Texte), Testset (900 Texte) Preprocessing unterzogen
- Beide Encodings werden in verschiedenen neuronalen Netzen trainiert
- Kombination mit bestem Ergebnis vergleichbare werden gekreuzt verglichen
- Vektorbasiertes Encoding ohne Netz
- Ordinal codierte Texte
- Vergleich Wörterbücher verschiedener Sparten

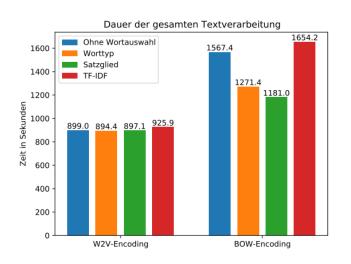


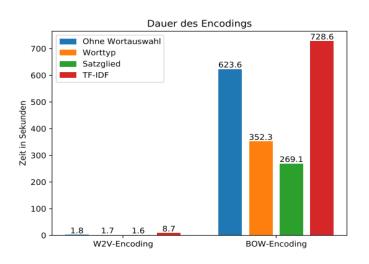
## Betrachtete Größen

- Trefferquoten
  - Anzahl korrekt zugeordneter Texte
  - Aufschlüsselung nach Sparte
- Bearbeitungszeit
  - Dauer der Preprocessing für Trainingsset
  - Rechtfertigt Trefferquote Dauer?



## Zeit







## Netzwerkausgaben

|              | Politik erkannt | Kultur erkannt | Gesellschaft erkannt | Leben erkannt | Sport erkannt | Reisen erkannt | Wirtschaft erkannt | Technik erkannt | Wissenschaft erkannt |
|--------------|-----------------|----------------|----------------------|---------------|---------------|----------------|--------------------|-----------------|----------------------|
| Politik      | 67              |                | 19                   |               | 2             |                | 10                 |                 |                      |
| Kultur       | 6               | 77             |                      |               |               | 6              | 4                  | 4               |                      |
| Gesellschaft | 9               | 5              | 64                   |               | 4             | 10             | 2                  | 3               | 2                    |
| Leben        | 2               | 4              | 4                    | 69            |               | 4              | 2                  | 2               | 13                   |
| Sport        |                 |                |                      |               | 89            | 5              | 3                  |                 | 0                    |
| Reisen       |                 |                |                      |               |               | 88             | 6                  | 5               | 0                    |
| Wirtschaft   | 3               |                |                      |               |               | 4              | 81                 | 6               | 4                    |
| Technik      |                 |                |                      |               |               | 14             | 6                  | 77              | 1                    |
| Wissenschaft |                 | 5              |                      | 26            |               | 5              | 4                  |                 | 58                   |

 BOW-codierte Texte, Wortauswahl nach Worttyp

|              | Politik erkannt | Kultur erkannt | Gesellschaft erkannt | Leben erkannt | Sport erkannt | Reisen erkannt | Wirtschaft erkannt | Technik erkannt | Wissenschaft erkannt |
|--------------|-----------------|----------------|----------------------|---------------|---------------|----------------|--------------------|-----------------|----------------------|
| Politik      | 84              | 3              | 8                    |               |               |                |                    |                 |                      |
| Kultur       | 7               | 70             | 11                   |               |               | 3              |                    | 2               | 4                    |
| Gesellschaft | 26              | 4              | 56                   |               |               |                | 6                  |                 | 3                    |
| Leben        | 7               | 2              | 11                   | 53            |               |                | 6                  | 2               | 19                   |
| Sport        |                 |                |                      | 3             | 79            | 3              | 6                  | 3               |                      |
| Reisen       | 1               |                | 4                    | 4             |               | 65             | 14                 | 3               | 7                    |
| Wirtschaft   | 18              |                |                      |               |               | 4              | 62                 | 9               | 6                    |
| Technik      | 2               |                |                      |               |               | 3              | 10                 | 68              | 15                   |
| Wissenschaft |                 |                | 6                    | 23            |               |                | 7                  |                 | 59                   |

W2V-codierte
Texte,
Wortauswahl
nach Worttyp



## Güte

|              | Ohne Wortauswahl | Worttyp | Satzglied | TF-IDF |
|--------------|------------------|---------|-----------|--------|
| W2V-Encoding | 63,11            | 66,22   | 60,78     | 62,78  |
| BOW-Encoding | 72,78            | 74,44   | 71,56     | 75,67  |

| • | Anteil korrekt |
|---|----------------|
|   | bestimmter     |
|   | Texte          |

|              | W2V-Encoding     |         |           |        | BOW-Encoding     |         |           |        |
|--------------|------------------|---------|-----------|--------|------------------|---------|-----------|--------|
|              | Ohne Wortauswahl | Worttyp | Satzglied | TF-IDF | Ohne Wortauswahl | Worttyp | Satzglied | TF-IDF |
| Politik      | 55               | 84      | 24        | 59     | 67               | 67      | 73        | 88     |
| Kultur       | 85               | 70      | 54        | 72     | 77               | 77      | 85        | 77     |
| Gesellschaft | 66               | 56      | 56        | 55     | 63               | 64      | 63        | 57     |
| Leben        | 71               | 53      | 36        | 64     | 83               | 69      | 62        | 73     |
| Sport        | 90               | 79      | 86        | 79     | 84               | 89      | 84        | 97     |
| Reisen       | 48               | 65      | 67        | 67     | 90               | 88      | 84        | 83     |
| Wirtschaft   | 36               | 62      | 63        | 48     | 76               | 81      | 69        | 69     |
| Technik      | 60               | 68      | 88        | 59     | 64               | 77      | 69        | 85     |
| Wissenschaft | 57               | 59      | 73        | 62     | 51               | 58      | 55        | 52     |

 Anteil korrekt bestimmter Texte pro Sparte



# Kreuzvergleich

|              | Politik erkannt | Kultur erkannt | Gesellschaft erkannt | Leben erkannt | Sport erkannt | Reisen erkannt | Wirtschaft erkannt | Technik erkannt | Wissenschaft erkannt |
|--------------|-----------------|----------------|----------------------|---------------|---------------|----------------|--------------------|-----------------|----------------------|
| Politik      | 69              |                |                      |               | 8             | 7              |                    | 16              | 0                    |
| Kultur       | 11              | 4              |                      |               | 36            | 12             |                    | 37              | 0                    |
| Gesellschaft | 17              |                | 17                   |               | 16            | 21             |                    | 29              | 0                    |
| Leben        |                 |                |                      | 39            | 8             | 10             |                    | 43              | 0                    |
| Sport        |                 |                |                      |               | 98            |                |                    |                 | 0                    |
| Reisen       |                 |                |                      |               | 5             | 80             |                    | 14              | 0                    |
| Wirtschaft   | 7               |                |                      |               | 12            | 6              | 15                 | 60              | 0                    |
| Technik      |                 |                |                      |               |               | 2              |                    | 97              | 0                    |
| Wissenschaft |                 |                |                      | 9             | 15            | 19             |                    | 55              | 2                    |

BOW-codierte
 Texte in W2V-Netz
 TF-IDF
 46.8%

|              | Politik erkannt | Kultur erkannt | Gesellschaft erkannt | Leben erkannt | Sport erkannt | Reisen erkannt | Wirtschaft erkannt | Technik erkannt | Wissenschaft erkannt |
|--------------|-----------------|----------------|----------------------|---------------|---------------|----------------|--------------------|-----------------|----------------------|
| Politik      | 60              | 7              | 18                   |               | 4             |                | 9                  |                 |                      |
| Kultur       | 6               | 82             | 3                    |               |               |                | 3                  | 3               |                      |
| Gesellschaft | 10              | 9              | 57                   |               | 5             | 10             | 3                  |                 | 3                    |
| Leben        |                 |                | 3                    | 58            | 2             | 5              |                    | 2               | 24                   |
| Sport        |                 | 6              | 2                    |               | 73            | 6              | 6                  | 6               |                      |
| Reisen       |                 |                | 2                    | 6             |               | 71             | 5                  | 7               | 6                    |
| Wirtschaft   | 10              | 5              |                      | 3             |               | 6              | 61                 | 7               | 4                    |
| Technik      | 2               | 12             |                      |               |               | 18             | 10                 | 53              | 4                    |
| Wissenschaft |                 |                | 2                    | 23            | 2             | 9              | 3                  | 1               | 58                   |

 W2V-codierte Texte in BOW-Netz
 Worttyp
 63.7%



# **Scoring und Ordinal Encoding**

|              | Politik erkannt | Kultur erkannt | Gesellschaft erkannt | Leben erkannt | Sport erkannt | Reisen erkannt | Wirtschaft erkannt | Technik erkannt | Wissenschaft erkannt |
|--------------|-----------------|----------------|----------------------|---------------|---------------|----------------|--------------------|-----------------|----------------------|
| Politik      | 50              | 5              | 2                    | 36            |               | 2              | 4                  | 1               | 0                    |
| Kultur       | 2               | 84             |                      | 6             |               | 3              | 2                  | 2               | 0                    |
| Gesellschaft | 8               | 9              | 26                   | 36            |               | 14             |                    |                 | 3                    |
| Leben        |                 |                |                      | 94            |               |                |                    |                 | 2                    |
| Sport        | 1               | 5              |                      | 4             | 70            | 14             | 4                  | 2               | 0                    |
| Reisen       |                 | 2              |                      | 11            |               | 75             | 7                  | 5               | 0                    |
| Wirtschaft   | 3               | 3              | 1                    | 7             |               | 9              | 70                 | 5               | 2                    |
| Technik      |                 |                |                      | 10            |               | 33             | 9                  | 46              | 0                    |
| Wissenschaft |                 |                |                      | 57            |               | 4              | 3                  |                 | 35                   |

Scoring Worttyp 61.1%

|              | Politik erkannt | Kultur erkannt | Gesellschaft erkannt | Leben erkannt | Sport erkannt | Reisen erkannt | Wirtschaft erkannt | Technik erkannt | Wissenschaft erkannt |
|--------------|-----------------|----------------|----------------------|---------------|---------------|----------------|--------------------|-----------------|----------------------|
| Politik      | 17              | 18             | 23                   | 4             | 8             | 11             | 8                  | 4               | 7                    |
| Kultur       | 23              | 17             | 10                   | 9             | 6             | 10             | 13                 | 9               | 3                    |
| Gesellschaft | 22              | 18             | 42                   | 4             | 2             | 5              | 3                  | 1               | 3                    |
| Leben        | 20              | 13             | 12                   | 14            | 9             | 9              | 9                  | 7               | 7                    |
| Sport        | 9               | 7              | 2                    | 13            | 12            | 13             | 13                 | 14              | 17                   |
| Reisen       | 10              | 15             | 5                    | 11            | 10            | 15             | 12                 | 13              | 9                    |
| Wirtschaft   | 8               | 9              | 4                    | 16            | 7             | 20             | 14                 | 8               | 14                   |
| Technik      |                 | 10             | 3                    | 11            | 17            | 18             | 15                 | 16              | 9                    |
| Wissenschaft | 10              | 18             | 4                    | 14            | 13            | 9              | 11                 | 12              | 9                    |

• Ordinal Worttyp 17.3%



## Fazit: Wortauswahlverfahren

- Wortauswahlverfahren kann Bearbeitungszeit verringern
- Auswahl nach Worttyp verbessert Ergebnis
- Auswahl nach Satzglied verringert Trefferquote
- TF-IDF wirkt sich unterschiedlich Encodings auf



# **Fazit: Encoding**

- Normalized BOW erzeugt h\u00f6here Trefferquoten
- Verringerung der Bearbeitungszeit durch Wortauswahl für BOW-Encoding empfohlen
- W2V-Encoding entnimmt Informationen aus Vorklassifizierung
- Positionierung von Begriffen keine relevante Auswirkung
- W2V-Encoding erzeugt bereits ohne Netz korrekte Ergebnisse



## **Ausblick**

- Weitere Parameter testen
  - Andere Netzart, Encoding
  - Vergleich verschiedener TF-IDF Grenzwerte
- Sortieren von Zeitungsarchiven
- Kombination mit Sentimentanalyse → Meinung von Autor über Thema



## Vielen Dank für Ihre Aufmerksamkeit!

Ihre Fragen



## Quellen

- 1) https://datasolut.com/textklassifikation/ (Acessed 1.8.2021)
- https://machinelearningknowledge.ai/data-preprocessing-in-machin e-learning/ (Acessed 1.8.2021)

## tu-freiberg.de

TU Bergakademie Freiberg 🔘 bergakademie\_freiberg 💟 TUBergakademie 🕒 TUBergakademie





#### **IMPRESSUM**

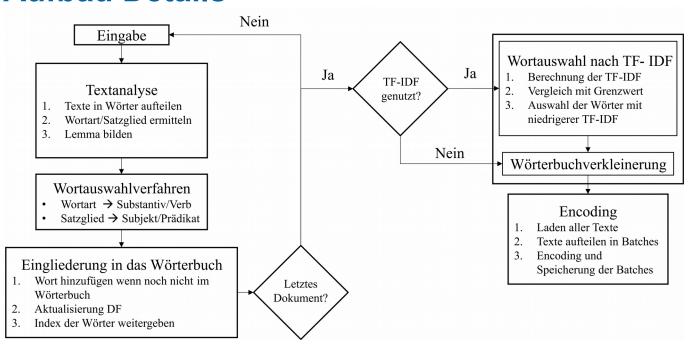
TU BERGAKADEMIE FREIBERG Universitätskommunikation Prüferstr. 2 09599 Freiberg Tel. +49(0)3731 39-2711, -3461





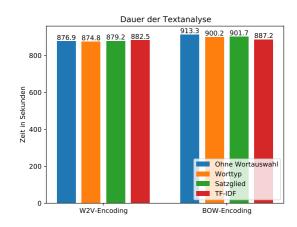


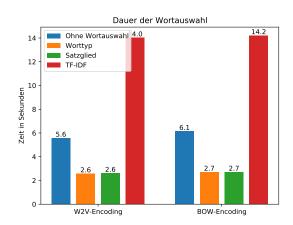
## **Aufbau Details**

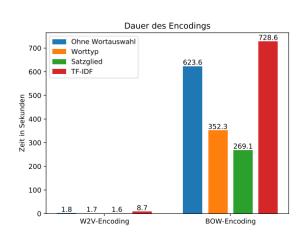




## Zeiten









## Güte

|              | W2V-Encoding     |         |           |                 | BOW-Encoding     |         |           |                 |
|--------------|------------------|---------|-----------|-----------------|------------------|---------|-----------|-----------------|
|              | Ohne Wortauswahl | Worttyp | Satzglied | TF- <b>I</b> DF | Ohne Wortauswahl | Worttyp | Satzglied | TF <b>-I</b> DF |
| Politik      | 20               | 63      | 8         | 34              | 15               | 21      | 21        | 37              |
| Kultur       | 44               | 14      | 14        | 30              | 14               | 16      | 44        | 34              |
| Gesellschaft | 38               | 42      | 51        | 34              | 18               | 27      | 30        | 21              |
| Leben        | 61               | 31      | 10        | 46              | 45               | 30      | 34        | 32              |
| Sport        | 14               | 7       | 41        | 7               | 7                | 7       | 13        | 8               |
| Reisen       | 13               | 16      | 32        | 62              | 88               | 48      | 42        | 26              |
| Wirtschaft   | 27               | 52      | 32        | 43              | 28               | 37      | 24        | 10              |
| Technik      | 9                | 23      | 105       | 36              | 15               | 22      | 19        | 27              |
| Wissenschaft | 106              | 56      | 60        | 43              | 15               | 22      | 29        | 24              |

 Falschzuordnungen der Sparte



## Güte

|              | W2V-Encoding          |                       |              |              | BOW-Encoding     |              |                       |                       |
|--------------|-----------------------|-----------------------|--------------|--------------|------------------|--------------|-----------------------|-----------------------|
|              | Ohne Wortauswahl      | Worttyp               | Satzglied    | TF-IDF       | Ohne Wortauswahl | Worttyp      | Satzglied             | TF- <b>I</b> DF       |
| Politik      | Gesel <b>l</b> schaft | Gese <b>ll</b> schaft | Gesellschaft | Gesellschaft | Gesellschaft     | Gesellschaft | Gese <b>ll</b> schaft | Gese <b>ll</b> schaft |
| Kultur       | Leben                 | Gese <b>ll</b> schaft | Technik      | Reisen       | Reisen           | Politik      | Po <b>l</b> itik      | Politik               |
| Gesellschaft | Kultur                | Politik               | Sport        | Politik      | Reisen           | Reisen       | Kultur                | Politik               |
| Leben        | Wissenschaft          | Wissenschaft          | Wissenschaft | Wissenschaft | Wissenschaft     | Wissenschaft | Wissenschaft          | Kultur                |
| Sport        | Kultur                | Wirtschaft            | Technik      | Reisen       | Reisen           | Reisen       | Kultur                | Reisen                |
| Reisen       | Wissenschaft          | Wirtschaft            | Wirtschaft   | Wirtschaft   | Wirtschaft       | Wirtschaft   | Kultur                | Technik               |
| Wirtschaft   | Wissenschaft          | Politik               | Technik      | Reisen       | Reisen           | Technik      | Technik               | Technik               |
| Technik      | Wissenschaft          | Wissenschaft          | Reisen       | Reisen       | Reisen           | Reisen       | Reisen                | Reisen                |
| Wissenschaft | Leben                 | Leben                 | Technik      | Leben        | Leben            | Leben        | Leben                 | Leben                 |

|              | W2V-Encoding     |         |           |        | BOW-Encoding     |         |           |        |  |
|--------------|------------------|---------|-----------|--------|------------------|---------|-----------|--------|--|
|              | Ohne Wortauswahl | Worttyp | Satzglied | TF-IDF | Ohne Wortauswahl | Worttyp | Satzglied | TF-IDF |  |
| Politik      | 20               | 8       | 31        | 17     | 14               | 19      | 13        | 9      |  |
| Kultur       | 5                | 11      | 20        | 8      | 7                | 6       | 5         | 9      |  |
| Gesellschaft | 9                | 26      | 12        | 17     | 12               | 10      | 15        | 17     |  |
| Leben        | 17               | 19      | 37        | 17     | 6                | 13      | 20        | 9      |  |
| Sport        | 3                | 6       | 6         | 10     | 12               | 5       | 6         | 2      |  |
| Reisen       | 18               | 14      | 9         | 11     | 4                | 6       | 3         | 6      |  |
| Wirtschaft   | 38               | 18      | 19        | 21     | 10               | 6       | 10        | 12     |  |
| Technik      | 16               | 15      | 5         | 11     | 26               | 14      | 13        | 5      |  |
| Wissenschaft | 31               | 23      | 9         | 26     | 34               | 26      | 28        | 28     |  |

# Spikes Falschzuordnungen

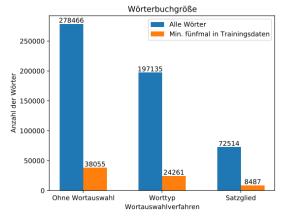
TU Bergakademie Freiberg | Vortragender: Erik Zimmermann| Gegenüberstellung von Preprocessing Methoden der Textklassifizierung von Zeitungsmeldungen durch künstliche neuronale Netze | 4.8.2021



## Wörterbuchvergleich

|              | Politik | Kultur | Gesellschaft | Leben | Sport | Reisen | Wirtschaft | Technik | Wissenschaft |
|--------------|---------|--------|--------------|-------|-------|--------|------------|---------|--------------|
| Politik      | 0       | 605    | 690          | 498   | 519   | 484    | 617        | 529     | 482          |
| Kultur       | 0       | 0      | 665          | 539   | 565   | 572    | 581        | 557     | 529          |
| Gesellschaft | 0       | 0      | 0            | 556   | 555   | 570    | 588        | 547     | 540          |
| Leben        | 0       | 0      | 0            | 0     | 480   | 500    | 567        | 559     | 635          |
| Sport        | 0       | 0      | 0            | 0     | 0     | 503    | 536        | 486     | 452          |
| Reisen       | 0       | 0      | 0            | 0     | 0     | 0      | 527        | 489     | 537          |
| Wirtschaft   | 0       | 0      | 0            | 0     | 0     | 0      | 0          | 621     | 560          |
| Technik      | 0       | 0      | 0            | 0     | 0     | 0      | 0          | 0       | 567          |
| Wissenschaft | 0       | 0      | 0            | 0     | 0     | 0      | 0          | 0       | 0            |

 Falschzuordnungen zu Sparte



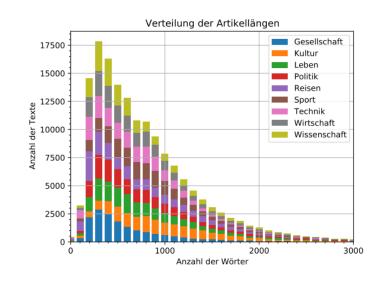
Wörterbuchgrößen

TU Bergakademie Freiberg | Vortragender: Erik Zimmermann| Gegenüberstellung von Preprocessing Methoden der Textklassifizierung von Zeitungsmeldungen durch künstliche neuronale Netze | 4.8.2021



# Eingangsdaten

- Daten aus verschiedenen
   Onlinepräsenzen von deutschen
   Zeitungen
- Neun Sparten angelehnt an Aufteilung der Zeitungen
- Trainingsset: 1000 Texte pro Sparte → 9000 Texte
- Testset: 100 Texte pro Sparte → 900 Texte
- Textlängen auf 1200 Wörter begrenzt





## **Probleme**

- Sparten nicht gleichartig verschieden
- Große Datenmengen in BOW-Encoding
  - → Begriffe min. fünf Mal in Trainingsdaten
  - → Encoding in Batches (50 Texte)
- Darstellung des BOW-Encoding abhängig von Textlänge
  - → Normalisierung von Encoding
  - → Nutzung verschiedener Netze
- W2V-Encoding keine konstante Länge
  - → Festlegung Maximallänge auf 1200 Wörter



## **TF-IDF**

Termfrequency

$$\text{TF-IDF}(w, T) = \frac{\text{Anzahl}(w, T)}{|T|} \cdot \log \left( \frac{|K|}{\text{Auswahl}(w, K)} \right)$$

- Vorkommenszahl von Wort in Text
- Je häufiger Wort in Text → relevanter für Text
- Inverse Documentfrequency
  - Anzahl Texte, die Begriff enthalten
  - Je mehr Texte Begriff enthalten → weniger relevant f
    ür Klassifizierung
- Eigener Grenzwert: 0.0019
- Termfrequency: 0.002 (1/500)
- Inverse Documentfrequency: 0.95 (log(1/9))



## **Convolution Neural Networks**

- Genutzt u.A. in Bilderkennung
- Mustererkennung unabhängig von Position in Daten

|   |   | l. |   |   |
|---|---|----|---|---|
| 0 | 1 | 1  | 1 | 0 |
| 1 | 0 | 1  | 0 | 0 |
| 1 | 1 | 1  | 1 | 1 |
| 0 | 1 | 1  | 0 | 0 |
| 1 | 0 | 0  | 1 | 1 |

- Convolution verrechnet Matrizen mit Teil der Eingangsdaten verrechnet
- Pooling fasst errechnete Daten verlustarm zusammen
- Benötigt Input mit konstanter Dimensionierung



## **Verwendete Netze**

#### **BOW**

#### Convolution Layer

- Filterzahl: 128
- Fenstergröße: 20
- Erweiterungsrate: 1
- Aktivierungsfunktion: Sigmoid
- Padding: Valid

#### MaxPooling Layer

- 1D
- Fenstergröße: 3

#### Batchnormalization

#### Flatten

#### Dense Layer

- Units: 9
- Aktivierungsfunktion: Softmax

#### W2V

#### Convolution Layer

- Filterzahl: 64
- Fenstergröße: 36
- Erweiterungsrate: 9
- Aktivierungsfunktion: Relu
- Padding: Valid

#### MaxPooling Layer

- 1D
- Fenstergröße: 2

#### Batchnormalization

#### Flatten

#### Dense Layer

- Units: 9
- Aktivierungsfunktion: Softmax

#### Ordinal

#### Convolution Layer

- Filterzahl: 64
- Fenstergröße: 4
- Erweiterungsrate: 1
- Aktivierungsfunktion:Relu
- Padding: Valid

#### MaxPooling Layer

- 1D
- Fenstergröße: 2

#### Batchnormalization

#### Flatten

#### Dense Layer

- Units: 9
- Aktivierungsfunktion:Softmax



## Wortmodifikationen

- Unterschiedlich deklinierte Begriffe zusammenfassen
- Lemmatisierung
  - Ermittelung Grundform
  - Berücksichtigung v. mehr Parametern
- Stemming
  - Bestimmung Wortstamm

| Startwort | Lemma | Stemm |
|-----------|-------|-------|
| die       | der   | die   |
| Bauern    | Bauer | Bau   |
| gingen    | gehen | ging  |
| um        | um    | um    |
| Häuser    | Haus  | Haus  |
| zu        | zu    | zu    |
| bauen     | bauen | bau   |



### **Ordinal Encoding**

|              | Politik erkannt | Kultur erkannt | Gesellschaft erkannt | Leben erkannt | Sport erkannt | Reisen erkannt | Wirtschaft erkannt | Technik erkannt | Wissenschaft erkannt |
|--------------|-----------------|----------------|----------------------|---------------|---------------|----------------|--------------------|-----------------|----------------------|
| Politik      | 26              | 16             | 28                   | 9             | 5             | 5              | 3                  |                 | 7                    |
| Kultur       | 13              | 17             | 11                   | 9             | 6             | 15             | 9                  | 10              | 10                   |
| Gesellschaft | 22              | 13             | 41                   | 7             | 3             |                | 7                  |                 | 4                    |
| Leben        | 19              | 12             | 13                   | 9             | 4             | 15             | 11                 | 8               | 9                    |
| Sport        | 11              | 8              | 3                    | 10            | 17            | 9              | 11                 | 15              | 16                   |
| Reisen       | 18              | 13             | 7                    | 8             | 5             | 10             | 13                 | 14              | 12                   |
| Wirtschaft   | 10              | 10             |                      | 8             | 10            | 16             | 18                 | 12              | 13                   |
| Technik      | 3               | 7              |                      | 16            | 16            | 13             | 15                 | 17              | 12                   |
| Wissenschaft | 14              | 11             | 6                    | 13            | 8             | 16             | 8                  | 15              | 9                    |

|  | Politik erkannt           | Kultur erkannt           | Gesellschaft erkannt       | Leben erkannt               | Sport erkannt                 | Reisen erkannt           | Wirtschaft erkannt      | lechnik erkannt           | wissenschaft erkannt   |
|--|---------------------------|--------------------------|----------------------------|-----------------------------|-------------------------------|--------------------------|-------------------------|---------------------------|------------------------|
| Politik                                  | 17                        | 18                       | 23                         | 4                           | 8                             | 11                       | 8                       | 4                         | 7                      |
| Kultur                                   | 23                        | 17                       | 10                         | 9                           | 6                             | 10                       | 13                      | 9                         | 3                      |
| Gesellschaft                             | 22                        | 18                       | 42                         | 4                           |                               | 5                        |                         |                           | 3                      |
| Leben                                    | 20                        | 13                       | 12                         | 14                          | 9                             | 9                        | 9                       | 7                         | 7                      |
| Sport                                    | 9                         | 7                        | 2                          | 13                          | 12                            | 13                       | 13                      | 14                        | 17                     |
| Reisen                                   | 10                        | 15                       | 5                          | 11                          | 10                            | 15                       | 12                      | 13                        | 9                      |
| Wirtschaft                               | 8                         | 9                        | 4                          | 16                          |                               | 20                       | 14                      | 8                         | 14                     |
| Technik                                  | 1                         | 10                       |                            | 11                          | 17                            | 18                       | 15                      | 16                        | 9                      |
| Wissenschaft                             | 10                        | 18                       | 4                          | 14                          | 13                            | 9                        | 11                      | 12                        | 9                      |
| Г  |                           |                          |                            |                             |                               |                          |                         |                           |                        |
|  | Politik erkannt           | Kultur erkannt           | Gesellschaft erkannt       | Leben erkannt               | Sport erkannt                 | Reisen erkannt           | Wirtschaft erkannt      | Technik erkannt           | Wissenschaft erkannt   |
| Politik                                  | Politik erkannt<br>15     | Kultur erkannt<br>9      | Gesellschaft erkannt<br>20 | Leben erkannt<br>8          | Sport erkannt<br>8            | Reisen erkannt<br>14     | Wirtschaft erkannt      | Technik erkannt           | Wissenschaft erkannt   |
| Politik<br>Kultur                        |                           |                          |                            |                             |                               |                          |                         |                           |                        |
|  | 15                        | 9                        | 20                         | 8                           | 8                             | 14                       | 5                       | 15                        | 6                      |
| Kultur                                   | 15<br>10                  | 9 23                     | 20<br>7                    | 10                          | 8                             | 14<br>13                 | 5<br>7                  | 15<br>12                  | 6<br>8                 |
| Kultur<br>Gesellschaft                   | 15<br>10<br>21            | 9<br>23<br>10            | 20<br>7<br>34              | 8<br>10<br>8                | 8<br>10<br>5                  | 14<br>13<br>7            | 5<br>7<br>5             | 15<br>12<br>5             | 6<br>8<br>5            |
| Kultur<br>Gesellschaft<br>Leben          | 15<br>10<br>21<br>18      | 9<br>23<br>10<br>10      | 20<br>7<br>34<br>5         | 8<br>10<br>8<br>7           | 8<br>10<br>5                  | 14<br>13<br>7<br>8       | 5<br>7<br>5             | 15<br>12<br>5<br>16       | 6<br>8<br>5            |
| Kultur<br>Gesellschaft<br>Leben<br>Sport | 15<br>10<br>21<br>18<br>8 | 9<br>23<br>10<br>10      | 20<br>7<br>34<br>5         | 8<br>10<br>8<br>7<br>7      | 8<br>10<br>5<br>12<br>22      | 14<br>13<br>7<br>6       | 5<br>7<br>5<br>10       | 15<br>12<br>5<br>16<br>22 | 6<br>8<br>5<br>14<br>9 |
| Kultur Gesellschaft Leben Sport Reisen   | 15<br>10<br>21<br>18<br>8 | 9<br>23<br>10<br>10<br>3 | 20<br>7<br>34<br>5<br>6    | 8<br>10<br>8<br>7<br>7<br>5 | 8<br>10<br>5<br>12<br>22<br>8 | 14<br>13<br>7<br>8<br>11 | 5<br>7<br>5<br>10<br>12 | 15<br>12<br>5<br>16<br>22 | 6<br>8<br>5<br>14<br>9 |

|              | Politik erkannt | Kultur erkannt | Gesellschaft erkannt | Leben erkannt | Sport erkannt | Reisen erkannt | Wirtschaft erkannt | Technik erkannt | Wissenschaft erkannt |
|--------------|-----------------|----------------|----------------------|---------------|---------------|----------------|--------------------|-----------------|----------------------|
| Politik      | 25              | 15             | 23                   | 5             | 6             | 10             | 7                  | 4               | 5                    |
| Kultur       | 12              | 19             | 8                    | 10            | 5             | 17             | 13                 | 10              | 6                    |
| Gesellschaft | 30              | 12             | 37                   | 5             | 3             | 3              | 4                  |                 | 5                    |
| Leben        | 26              | 14             | 6                    | 8             |               | 19             | 10                 | 8               | 6                    |
| Sport        | 6               | 9              | 3                    | 19            | 15            | 8              | 14                 | 14              | 12                   |
| Reisen       | 16              | 12             | 6                    | 12            | 5             | 10             | 15                 | 9               | 15                   |
| Wirtschaft   | 5               | 13             | 3                    | 7             | 7             | 18             | 16                 | 15              | 16                   |
| Technik      | 4               | 11             | 2                    | 14            | 11            | 12             | 14                 | 16              | 16                   |
| Wissenschaft | 15              | 8              |                      | 19            | 9             | 16             | 7                  | 14              | 9                    |

Ohne Wortauswahl18.2%

Worttyp17.3%

Satzglied17.9%

• TF-IDF 17.2%

Preprocessing Methoden der



### **Scoring Encoding**

|                                 | Politik erkannt | Kultur erkannt | Gesellschaft erkannt | Leben erkannt  | Sport erkannt | Reisen erkannt   | Wirtschaft erkannt | Technik erkannt | Wissenschaft erkannt |
|---------------------------------|-----------------|----------------|----------------------|----------------|---------------|------------------|--------------------|-----------------|----------------------|
| Politik                         | 51              | 4              | 0                    | 32             | 0             | 7                | 5                  | 1               | 0                    |
| Kultur                          | 7               | 75             | 0                    | 3              | 1             | 10               | 2                  |                 | 0                    |
| Gesellschaft                    | 11              | 8              | 21                   | 34             | 4             | 18               | 2                  |                 | 2                    |
| Leben                           |                 |                | 0                    | 94             | 0             | 1                |                    |                 | 3                    |
| Sport                           |                 |                |                      | 3              | 69            | 25               |                    |                 | 0                    |
| Reisen                          |                 |                |                      | 13             | 0             | 78               | 5                  | 3               | 0                    |
| Wirtschaft                      |                 |                |                      | 14             |               | 29               | 51                 |                 | 2                    |
| Technik                         |                 |                |                      | 11             |               | 50               | 4                  | 33              | 1                    |
| Wissenschaft                    |                 |                |                      | 51             | 0             | 10               |                    |                 | 35                   |
|                                 | Politik erkannt | Kultur erkannt | Gesellschaft erkannt | Leben erkannt  | Sport erkannt | Reisen erkannt   | Wirtschaft erkannt | Technik erkannt | Wissenschaft erkannt |
| Politik                         | 50              | 5              | 2                    | 36             | 0             | 2                | 4                  | 1               | 0                    |
| Kultur                          | 2               | 84             | 0                    | 6              | 1             | 3                |                    |                 | 0                    |
| Gesellschaft                    | 8               | 9              | 26                   | 36             | 2             | 14               | 2                  |                 | 3                    |
| Leben                           | 1               | 1              | 0                    | 94             | 0             | 1                |                    |                 | 2                    |
| Sport                           |                 | 5              |                      | 4              | 70            | 14               | 4                  |                 | 0                    |
| Reisen                          |                 | 2              |                      | 11             | 0             | 75               | 7                  | 5               | 0                    |
| Wirtschaft                      | 3               | 3              |                      | 7              |               | 9                | 70                 | 5               | 2                    |
| Technik                         |                 |                |                      | 10             | 1             | 33               | 9                  | 46              | 0                    |
| Wissenschaft                    |                 |                |                      | 57             |               | 4                |                    |                 | 35                   |
| Г                               | Politik erkannt | Kultur erkannt | Gesellschaft erkannt | Leben erkannt  | Sport erkannt | Reisen erkannt   | Wirtschaft erkannt | Technik erkannt | Wissenschaft erkannt |
| Politik                         | 37              | 4              | Gesellschaft erkannt | 50             | o<br>0        | Reisell erkallit | Wirtschaft erkannt | 3               | Wissenschaft erkannt |
| Kultur                          | 2               | 55             | 0                    | 23             | 1             | 16               | 1                  | 1               | 1                    |
| Gesellschaft                    | 6               | 4              | 30                   | 42             | 2             | 15               | 0                  | 0               | 1                    |
| Leben                           | 1               |                | 0                    | 79             | 0             | 3                |                    |                 | 16                   |
| Sport                           | 2               | 4              | 2                    | 20             | 45            | 18               |                    | 5               | Δ                    |
| Reisen                          | 0               | 1              | 0                    | 23             | 0             | 66               | 2                  | 5               | 3                    |
| Wirtschaft                      | 7               |                | 1                    | 40             | 0             | 12               | 19                 | 16              | 5                    |
| Technik                         | 2               | 3              | 1                    | 17             | 1             | 37               | 2                  | 36              | 1                    |
| Wissenschaft                    | 2               |                | 0                    | 47             | 0             | 4                | 0                  | 3               | 44                   |
|                                 |                 |                |                      |                |               |                  |                    |                 |                      |
|                                 | Politik erkannt | Kultur erkannt | Gesellschaft erkannt | Leben erkannt  | Sport erkannt | Reisen erkannt   | Wirtschaft erkannt | Technik erkannt | Wissenschaft erkannt |
| Politik                         | 54              | 3              | 0                    | 32             | 0             | 5                | 5                  | 1               | 0                    |
| Kultur                          | 7               | 75             | 0                    | 3              | 1             | 10               | 2                  |                 | 0                    |
| Gesellschaft                    | 11              | 8              | 21                   | 35             | 4             | 17               | 2                  |                 | 2                    |
| Leben                           | 1               |                | 0                    | 94             | 0             | 1                |                    |                 | 3                    |
| Sport                           | 0               |                | 0                    | 4              | 69            | 24               | 1                  | 0               | 0                    |
| Reisen                          |                 |                |                      | 12             |               | 79               | 5                  | 3               | 0                    |
|                                 |                 |                |                      |                |               | 28               | 53                 |                 | 2                    |
| Wirtschaft                      | 0               |                | 1                    | 13             |               |                  |                    | 3               |                      |
| Wirtschaft Technik Wissenschaft | 0 0             |                | 0 0                  | 13<br>10<br>50 | 0             | 45<br>10         | 4                  | 39              | 1 35                 |

Ohne Wortauswahl 56.3%

• Worttyp 61.1%

Satzglied45.7%

• TF-IDF 57.7%

Preprocessing Methoden der



### Vektorrepräsentation

|   | Politik erkannt   | Kultur erkannt  | Gesellschaft erkannt  | Leben erkannt   | Sport erkannt   | Reisen erkannt   | Wirtschaft erkannt                                       | Technik erkannt                                    | Wissenschaft erkannt   |
|---|---|---|---|---|---|--|--|--|--|
| Politik   | 55  | 12  | 20  |   | 1   |  |  |  | 7  |
| Kultur  | 4   | 85  | 2   | 5   | 0   |  | 3  |  | 1  |
| Gesellschaft  | 5   | 9   | 66  | 5   | 1   | 4  | 3  |  | 7  |
| Leben   |   | 3   | 4   | 71  | 0   |  |  |  | 17   |
| Sport   |   | 3   | 2   |   | 90  | 1  |  |  | 2  |
| Reisen  |   | 7   | 5   | 6   | 4   | 48   | 6  | 5  | 18   |
| Wirtschaft  | 6   |   | 3   | 5   | 5   | 3  | 36   | 3  | 38   |
| Technik   |   | 6   | 1   | 6   | 0   | 5  | 5  | 60   | 16   |
| Wissenschaft  |   | 3   |   | 31  | 3   |  | 5  |  | 57   |
|   | Politik erkannt   | Kultur erkannt  | Gesellschaft erkannt  | Leben erkannt   | Sport erkannt   | Reisen erkannt   | Wirtschaft erkannt                                       | Technik erkannt                                    | Wissenschaft erkannt   |
| Politik   | 84  | 3   | 8   | 1   | 1   | 0  | 1  | 1  | 1  |
| Kultur  | 7   | 70  | 11  |   |   | 3  | 2  |  | 4  |
| Gesellschaft  | 26  | 4   | 56  |   | 2   | 2  | 6  |  | 3  |
| Leben   | 7   | 2   | 11  | 53  | 0   | 0  | 6  |  | 19   |
| Sport   | 2   |   | 1   | 3   | 79  | 3  | 6  | 3  | 1  |
| Reisen  |   |   | 4   | 4   | 1   | 65   | 14   | 3  | 7  |
| Wirtschaft  | 18  |   | 0   |   | 0   | 4  | 62   | 9  | 6  |
| Technik   | 2   |   | 1   |   |   | 3  | 10   | 68   | 15   |
| Wissenschaft  |   |   | 6   | 23  | 2   |  | 7  | 2  | 59   |
|   |   |   |   |   |   |  |  |  |  |
|   | Politik erkannt   | Kultur erkannt  | Gesellschaft erkannt  | Leben erkannt   | Sport erkannt   | Reisen erkannt   | Wirtschaft erkannt                                       | Technik erkannt                                    | Wissenschaft erkannt   |
| Politik   | Politik erkannt<br>24   | Kultur erkannt<br>4   | Gesellschaft erkannt<br>31  | Leben erkannt   | Sport erkannt   | Reisen erkannt<br>0  | Wirtschaft erkannt                                       | Technik erkannt<br>24                              | Wissenschaft erkannt   |
| Politik<br>Kultur   |   |   |   |   |   |  |  |  |  |
|   | 24  | 4   | 31  | 0   | 4   | 0  | 10   | 24   | 3  |
| Kultur  | 24  | 4<br>54   | 31  | 0   | 4   | 0 4  | 10<br>3  | 24   | 3  |
| Kultur<br>Gesellschaft  | 24<br>1<br>4  | 4<br>54<br>4  | 31<br>1<br>56   | 0 0 1   | 4<br>17<br>12   | 0<br>4<br>8  | 10<br>3<br>3   | 24<br>20<br>6                                      | 3<br>0<br>6  |
| Kultur<br>Gesellschaft<br>Leben   | 24<br>1<br>4  | 4<br>54<br>4<br>1   | 31<br>1<br>56<br>8  | 0<br>0<br>1<br>36                                     | 4<br>17<br>12<br>0  | 0<br>4<br>8<br>3   | 10<br>3<br>3   | 24<br>20<br>6<br>13                                | 3<br>0<br>6<br>37  |
| Kultur<br>Gesellschaft<br>Leben<br>Sport  | 24<br>1<br>4<br>1   | 4<br>54<br>4<br>1   | 31<br>1<br>56<br>8  | 0<br>0<br>1<br>36                                     | 4<br>17<br>12<br>0<br>86  | 0<br>4<br>8<br>3<br>2  | 10<br>3<br>3<br>1  | 24<br>20<br>6<br>13<br>6                           | 3<br>0<br>6<br>37<br>3   |
| Kultur<br>Gesellschaft<br>Leben<br>Sport<br>Reisen  | 24<br>1<br>4<br>1<br>0  | 4<br>54<br>4<br>1<br>1<br>3                                 | 31<br>1<br>56<br>8<br>1   | 0<br>0<br>1<br>36<br>0                                | 4<br>17<br>12<br>0<br>86  | 0<br>4<br>8<br>3<br>2  | 10<br>3<br>3<br>1<br>1                                   | 24<br>20<br>6<br>13<br>6                           | 3<br>0<br>6<br>37<br>3<br>7                                    |
| Kultur Gesellschaft Leben Sport Reisen Wirtschaft   | 24<br>1<br>4<br>1<br>0<br>0                                     | 4<br>54<br>4<br>1<br>1<br>3                                 | 31<br>1<br>56<br>8<br>1<br>3  | 0<br>0<br>1<br>36<br>0                                | 4<br>17<br>12<br>0<br>86<br>3   | 0<br>4<br>8<br>3<br>2<br>67  | 10<br>3<br>3<br>1<br>1<br>9                              | 24<br>20<br>6<br>13<br>6<br>8                      | 3<br>0<br>6<br>37<br>3<br>7                                    |
| Kultur Gesellschaft Leben Sport Reisen Wirtschaft Technik   | 24<br>1<br>4<br>1<br>0<br>0<br>2<br>0<br>0                      | 4<br>54<br>4<br>1<br>1<br>3<br>1<br>0                       | 31<br>1<br>56<br>8<br>1<br>3<br>4<br>1  | 0<br>0<br>1<br>36<br>0<br>0<br>1<br>0                 | 4<br>17<br>12<br>0<br>86<br>3<br>1<br>2   | 0<br>4<br>8<br>3<br>2<br>67<br>6<br>5  | 10<br>3<br>3<br>1<br>1<br>9<br>63                        | 24<br>20<br>6<br>13<br>6<br>8<br>19<br>88          | 3<br>0<br>6<br>37<br>3<br>7<br>3<br>1                          |
| Kultur Gesellschaft Leben Sport Reisen Wirtschaft Technik   | 24<br>1<br>4<br>1<br>0<br>0<br>2                                | 4<br>54<br>4<br>1<br>1<br>3<br>1                            | 31<br>1<br>56<br>8<br>1<br>3<br>4   | 0<br>0<br>1<br>36<br>0<br>0                           | 4<br>17<br>12<br>0<br>86<br>3<br>1  | 0<br>4<br>8<br>3<br>2<br>67<br>6<br>5  | 10<br>3<br>3<br>1<br>1<br>1<br>9<br>63<br>3              | 24<br>20<br>6<br>13<br>6<br>8<br>19                | 3<br>0<br>6<br>37<br>3<br>7<br>3                               |
| Kultur Gesellschaft Leben Sport Reisen Wirtschaft Technik Wissenschaft  | 24 1 1 0 0 2 0 0 Politik erkannt                                | 4 54 4 1 1 1 3 3 1 1 0 0 0 Kultur erkannt                   | 31<br>1<br>56<br>8<br>1<br>3<br>4<br>1<br>2<br>Gesellschaft erkannt                       | 0<br>0<br>1<br>36<br>0<br>0<br>1<br>0                 | 4 17 12 0 86 3 1 1 2 2 2 Sport erkannt  | 0 4 8 3 2 67 6 5 4 Reisen erkannt  | 10 3 3 3 1 1 1 1 9 63 3 3 7 7 Wirtschaft erkannt         | 24 20 6 13 6 8 19 88 19 7 Echnik erkannt           | 3 0 6 37 3 7 3 1 73 Wissenschaft erkannt                       |
| Kultur Gesellschaft Leben Sport Reisen Wirtschaft Technik Wissenschaft  | 24  1  4  1  0  0  2  0  0  Politik erkannt 59                  | 4 54 4 1 1 1 3 3 1 1 0 0 Waltur erkannt 10                  | 31<br>1<br>56<br>8<br>1<br>3<br>4<br>1<br>2<br>Cesellschaft erkannt                       | 0 0 1 36 0 1 1 0 8 Leben erkannt                      | 4 17 12 0 86 3 1 1 2 2 2 Sport erkannt  | 0 4 8 3 2 67 6 5 4 Reisen erkannt  | 10 3 3 1 1 9 63 3 2 Wirtschaft erkannt                   | 24 20 6 13 6 8 19 88 9 Technik erkannt             | 3 0 6 37 3 7 3 1 7 3 Wissenschaft erkannt:                     |
| Kultur Gesellschaft Leben Sport Reisen Wirtschaft Technik Wissenschaft Politik Kultur                                 | 24  1  4  1  0  0  2  0  Politik erkannt  59                    | 4 54 4 1 1 1 1 3 3 1 1 0 0 Wulkur erkannt 10 72             | 31<br>1<br>56<br>8<br>1<br>3<br>4<br>1<br>7<br>Gesellschaft erkannt                       | 0 0 1 36 0 1 0 1 0 Eben erkant 1 0                    | 4 17 12 0 86 3 1 1 2 2 2 Sport erkannt 2 1 1  | 0 4 6 3 2 67 6 5 4 Reisen erkannt 1  | 10 3 3 1 1 1 9 63 3 7 Wirtschaft erkannt 6 4             | 24 20 6 13 6 8 19 88 9 Technik erkannt 3           | 3 0 6 37 3 7 3 1 7 3 Wissenschaft erkannt 1 2                  |
| Kultur Gesellschaft Leben Sport Reisen Wirtschaft Technik Wissenschaft Politik Kultur Gesellschaft                    | 24 1 4 1 0 0 0 2 0 Politic erkannt 59 6 17                      | 4 54 4 1 1 1 3 3 1 1 0 0 0 Kultur erkannt 10 72 10 10       | 31<br>1<br>56<br>1<br>3<br>4<br>1<br>2<br>Cesellschaft erkannt                            | 0 0 1 36 0 0 1 0 1 Leben erkannt 1 0 3                | 4 17 12 0 86 3 1 1 2 2 2 Sport erkannt 2 1 3 3  | 0 4 6 3 2 67 6 5 5 4 4 Reisen erkannt. 1 8 4 4   | 10 3 3 1 1 1 9 63 3 2 Wirtschaft erkannt 4               | 24 20 6 13 6 8 19 88 19 Technik erkantt 4          | 3 0 6 37 3 7 3 1 1 77 77 73                                    |
| Kultur Gesellschaft Leben Sport Reisen Wirtschaft Technik Wissenschaft Vissenschaft Kultur Gesellschaft Leben         | 24  1  4  1  0  2  0  Politik erkannt  59  6  17                | 4 4 4 1 1 1 1 3 3 1 1 0 0 0 Kultur erkannt 10 72 10 10 1 1  | 31<br>1<br>56<br>3<br>4<br>1<br>2<br>Gesellschaft erkannt<br>17<br>3<br>55                | 0 0 1 36 0 0 1 1 0 0 1 Leben erkannt 1 0 3 64         | 4 17 12 0 86 3 1 1 2 2 2 2 Sport erkannt 2 1 3 0 0  | 0 4 6 3 7 67 6 5 4 Reisen erkannt 1 8 4  | 10 3 3 1 1 1 9 63 3 2 Wirtschaft erkannt 6 4 2 4         | 24 20 6 13 6 8 19 88 9 7echnik erkannt 3 4 2 2     | 3 6 37 1 7 3 1 1 73 Wissenschaft erkannt 1 2 4 17              |
| Kultur Gesellschaft Leben Sport Reisen Wirtschaft Technik Wissenschaft Kultur Gesellschaft Leben Sport                | 24 1 4 1 0 0 2 0 0 Politik erkannt 59 6 17 3                    | 4 4 54 4 1 1 1 3 3 1 1 0 0 0 Kultur erkannt 10 72 10 11 3 3 | 31<br>1<br>56<br>3<br>1<br>3<br>4<br>1<br>2<br>Ceselschaft erkannt<br>17<br>3<br>55       | 0 0 1 36 0 0 1 1 0 0 8 8 Leben erkannt 1 0 3 3 64 1 1 | 4 17 12 0 0 06 3 1 1 2 2 2 5 Sport erkannt 2 1 3 0 0 79                                   | 0 4 8 3 3 2 67 66 5 4 4 8 elsen erkannt 1 8 4 2 10 10 10 10 10 10 10 10 10 10 10 10 10 | 10 3 3 1 1 9 63 3 2 Wirtschaft erkannt 6 4 2             | 24 20 6 13 6 8 19 88 9 Technik erkannt 3 4 2 2 2   | 3 0 6 37 3 3 1 7 3 3 1 7 3 3 1 1 7 3 3 1 1 7 3 1 1 7 3 1 1 1 1 |
| Kultur Gesellschaft Leben Sport Reisen Wirtschaft Technik Wissenschaft Politik Kultur Gesellschaft Leben Sport Reisen | 24  1  2  1  0  0  2  0  0  Politik erkannt  59  6  17  3  0  3 | 4   | 31<br>1<br>56<br>8<br>1<br>3<br>4<br>1<br>2<br>Ceselischaft erkannt<br>17<br>3<br>35<br>7 | 0 0 1 1 36 0 0 1 1 0 8 Leben erfannt 1 0 3 64         | 4 17 17 12 0 06 3 1 1 2 2 2 Sport erkannt 2 1 1 3 0 0 79 0 0                              | 0 4 8 3 3 2 6 6 7 6 6 5 5 4 4 Reisen erkannt 1 8 4 2 10 6 7                            | 10 3 3 1 1 1 9 63 3 2 Wirtschaft erkannt 6 4 2 4 2 11    | 24 20 6 13 6 8 19 88 9 Technik erkannt 3 4 2 2 10  | 3 0 6 37 3 7 3 1 73 Wissenschaft erkannt 1 2 4 17 0            |
| Kultur Gesellschaft Leben Sport Reisen Wirtschaft Technik Wissenschaft Kultur Gesellschaft Leben Sport Reisen         | 24  1  2  0  0  2  0  0  Politic erkannt  59  6  17  3  4       | 4   | 31<br>1<br>56<br>8<br>1<br>1<br>2<br>Cesellschaft erkannt<br>17<br>3<br>55<br>7           | 0 0 1 36 0 0 1 1 0 8 Leben erkannt 1 0 3 64 1 2       | 4 17 12 0 0 06 3 1 1 2 2 2 5 Sport erkannt 2 1 1 3 0 0 79 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 | 0 4 6 3 3 2 67 6 6 5 4 4 4 2 2 1 10 67 21  | 10 3 3 1 1 1 9 63 3 2 Wirtschaft erkannt 4 2 4 4 2 11 48 | 24 20 6 13 6 8 19 88 19 Technik erkannt 3 4 2 2 10 | 3 0 6 37 3 7 3 1 7 3 1 2 4 17 0 4                              |

Ohne Wortauswahl 63.1%

Worttyp66.2%

Satzglied 60.8%

• TF-IDF 62.8%



#### **BOW Encoding**

|  | Politik erkannt  | Kultur erkannt  | Gesellschaft erkannt  | Leben erkannt                                  | Sport erkannt                                 | Reisen erkannt  | Wirtschaft erkannt  | Technik erkannt   | Wissenschaft erkannt   |
|--|--|---|---|--|---|---|---|---|--|
| Politik  | 67   | 1   | 14  | 2  | 2   | 4   | 10  | 0   | 0  |
| Kultur   | 6  | 77  | 0   |  | 1   | 7   | 2   | 5   | 1  |
| Gesellschaft   | 6  | 8   | 63  | 2  | 3   | 12  | 3   | 1   | 2  |
| Leben  | 0  | 3   | 2   | 83   | 0   | 5   | 1   |   | 6  |
| Sport  | 0  | 1   | 0   | 1  | 84  | 12  | 2   | 0   | 0  |
| Reisen   | 1  | 0   |   |  | 0   | 90  | 4   | 4   | 1  |
| Wirtschaft   | 2  |   |   | 2  | 0   | 10  | 76  | 5   | 4  |
| Technik  | 0  |   | 1   | 3  | 0   | 26  | 4   | 64  | 1  |
| Wissenschaft   | 0  |   | 0   | 34   | 1   | 12  | 2   | 0   | 51   |
|  | Politik erkannt  | Kultur erkannt  | Gesellschaft erkannt  | Leben erkannt                                  | Sport erkannt                                 | Reisen erkannt  | Wirtschaft erkannt  | Technik erkannt   | Wissenschaft erkann  |
| Politik  | 67   | 0   | 19  | 0  | 2   | 0   | 10  | 1   | Wissenschalt erkann  |
| Kultur   | 6  | 77  | 19  |  | 0   | 6   | 4   | 4   | 1  |
|  |  | 5   |   |  | -   | -   |   |   |  |
| Gesellschaft   | 9  | -   | 64  | 1  | 4   | 10  | 2   | 3   | 2  |
| Leben  |  | 4   | 4   | 69   | 0   | 4   | 2   |   | 13   |
| Sport  |  |   | 1   | 1  | 89  | 5   | 3   | 0   | 0  |
| Reisen   | 0  |   | 0   |  | 0   | 88  | 6   | 5   | 0  |
| Wirtschaft   | 3  |   | 1   | 1  | 0   | 4   | 81  | 6   | 4  |
| Technik  |  | 1   | 1   | 0  | 0   | 14  | 6   | 77  | 1  |
| Wissenschaft   | 0  | 5   | 0   | 26   | 1   | 5   | 4   | 1   | 58   |
|  | Politik erkannt  | Kultur erkannt  | Gesellschaft erkannt  | Leben erkannt                                  | Sport erkannt                                 | Reisen erkannt  | Wirtschaft erkannt  | Technik erkannt   | Wissenschaft erkann  |
|  |  |   |   |  |   |   |   |   | Wissensendre erkann  |
| Politik  | 73   | 5   | 13  | 1  | 3   | 0   | 4   | 1   | 0  |
| Politik<br>Kultur  | 73<br>5  | 5<br>85   | 13<br>2   | 0  | 3   |   |   | 2   |  |
|  |  |   |   | 0 3  | 3   | 0   | 4   | 1   | 0  |
| Kultur   | 5  | 85  | 2   |  | 3   | 3   | 2   | 2   | 0  |
| Kultur<br>Gesellschaft   | 5<br>7   | 85<br>15  | 2<br>63   | 3  | 3<br>1<br>3                                   | 0<br>3<br>6   | 2 2   | 2 0   | 0 0 1  |
| Kultur<br>Gesellschaft<br>Leben  | 5<br>7<br>2  | 85<br>15<br>5   | 2<br>63<br>6  | 3<br>62  | 3<br>1<br>3<br>0                              | 0<br>3<br>6<br>2  | 2 2 2   | 1<br>2<br>0   | 0<br>0<br>1<br>20  |
| Kultur<br>Gesellschaft<br>Leben<br>Sport   | 5<br>7<br>2<br>1   | 85<br>15<br>5   | 2<br>63<br>6  | 3<br>62<br>0                                   | 3<br>1<br>3<br>0                              | 0<br>3<br>6<br>2<br>5   | 4<br>2<br>2<br>2<br>2                                       | 1<br>2<br>0<br>1  | 0<br>0<br>1<br>20  |
| Kultur Gesellschaft Leben Sport Reisen   | 5<br>7<br>2<br>1<br>0  | 85<br>15<br>5<br>6  | 2<br>63<br>6<br>2   | 3<br>62<br>0                                   | 3<br>1<br>3<br>0<br>84                        | 0<br>3<br>6<br>2<br>5   | 4<br>2<br>2<br>2<br>2<br>2<br>3                             | 1<br>2<br>0<br>1<br>0<br>3                                | 0<br>0<br>1<br>20<br>0<br>3  |
| Kultur Gesellschaft Leben Sport Reisen Wirtschaft  | 5<br>7<br>2<br>1<br>0  | 85<br>15<br>5<br>6<br>3   | 2<br>63<br>6<br>2<br>2  | 3<br>62<br>0<br>0                              | 3<br>1<br>3<br>0<br>84<br>2                   | 0<br>3<br>6<br>2<br>5<br>84<br>6  | 2<br>2<br>2<br>2<br>2<br>3<br>69                            | 1<br>2<br>0<br>1<br>0<br>3                                | 0<br>0<br>1<br>20<br>0<br>3  |
| Kultur Gesellschaft Leben Sport Reisen Wirtschaft Technik  | 5<br>7<br>2<br>1<br>0<br>5   | 85<br>15<br>5<br>6<br>3<br>2  | 2<br>63<br>6<br>2<br>2<br>2                                       | 3<br>62<br>0<br>0<br>1                         | 3<br>1<br>3<br>0<br>84<br>2<br>2              | 0<br>3<br>6<br>2<br>5<br>84<br>6  | 4<br>2<br>2<br>2<br>2<br>2<br>3<br>69                       | 1<br>2<br>0<br>1<br>0<br>3<br>10                          | 0<br>0<br>1<br>20<br>0<br>3<br>3<br>3<br>2                         |
| Kultur Gesellschaft Leben Sport Reisen Wirtschaft Technik  | 5<br>7<br>2<br>1<br>0<br>5<br>0  | 85<br>15<br>5<br>6<br>3<br>2<br>5   | 2<br>63<br>6<br>2<br>2<br>2<br>2<br>3                             | 3<br>62<br>0<br>0<br>1<br>1<br>28              | 3<br>1<br>3<br>0<br>84<br>2<br>2<br>1         | 0<br>3<br>6<br>2<br>5<br>84<br>6<br>13                                  | 4<br>2<br>2<br>2<br>2<br>2<br>3<br>69<br>6                  | 1 2 0 1 1 0 3 10 69 2                                     | 0<br>0<br>1<br>20<br>0<br>3<br>3<br>3<br>2                         |
| Kultur Gesellschaft Leben Sport Reisen Wirtschaft Technik Wissenschaft   | 5 7 2 1 0 5 0 1 1 Politik erkannt  | 85<br>15<br>5<br>6<br>3<br>2<br>5<br>3<br>Kultur erkannt                                    | 2 63 6 2 2 2 2 3 3 0 9 Gesellschaft erkonnt                       | 3 62 0 0 1 1 1 28 Leben erkannt                | 3<br>1<br>3<br>0<br>84<br>2<br>2<br>1         | 0 3 6 2 5 84 6 13 7 7 Reisen erkannt                                    | 4<br>2<br>2<br>2<br>2<br>2<br>3<br>69<br>6                  | 1 2 0 1 1 0 3 3 10 69 2 Technik erkannt                   | 0<br>0<br>1<br>20<br>0<br>3<br>3                                   |
| Kultur Gesellschaft Leben Sport Reisen Wirtschaft Technik Wissenschaft   | 5 7 2 1 0 5 0 1 1 Politik erkannt 88   | 85<br>15<br>5<br>6<br>3<br>2<br>5<br>3<br>Kultur erkannt                                    | 2 63 6 2 2 2 3 0 0 Gesellschaft erkannt                           | 3 62 0 0 1 1 1 28 Leben erkannt 0              | 3 1 3 0 04 2 2 1 1 1 Sport erkannt 1          | 0<br>3<br>6<br>2<br>5<br>84<br>6<br>13<br>7                             | 4 2 2 2 2 2 3 69 6 3 3 Wirtschaft erkannt 1                 | 1 2 0 1 1 0 0 1 1 0 0 0 1 0 0 0 0 0 0 0                   | 0 0 1 20 0 3 3 3 2 55 5 Wissenschaft erkann                        |
| Kultur Gesellschaft Leben Sport Reisen Wirtschaft Technik Wissenschaft Politik Kultur                          | 5 7 2 1 0 5 0 1 Politik erkannt 88   | 85<br>15<br>5<br>6<br>3<br>2<br>5<br>3<br>Kultur erkannt<br>0                               | 2 63 6 2 2 2 3 0 Gesellschaft erkonnt                             | 3 62 0 0 0 1 1 1 28 Leben erkannt 0 0 0        | 3 1 3 0 84 2 2 1 1 Sport erkannt 1            | 0 3 6 2 5 84 6 13 7 Reisen erkannt 0                                    | 4 2 2 2 2 3 69 6 3 Wirtschaft erkannt                       | 1 2 0 1 1 0 0 3 10 0 69 2 Technik erkannt 0 5             | 0 0 1 20 0 3 3 3 2 55 Wissenschaft erkann 1 1                      |
| Kultur Gesellschaft Leben Sport Reisen Wirtschaft Technik Wissenschaft Politik Kultur Gesellschaft Leben       | 5 7 2 1 0 5 0 1 Politik erfannt 88 9   | 85<br>15<br>5<br>6<br>3<br>2<br>5<br>3<br>Kultur erkannt<br>0                               | 2 63 6 7 2 2 2 3 3 9 9 Gesellschaft erkonnt 9 4 4 57              | 3 62 0 0 0 1 1 28 Leben erkannt 0 0 1 1        | 3 1 3 0 84 2 2 1 1 Sport erkannt 1 1 3        | 0 3 6 2 5 84 6 13 7 Reisen erkannt 0 2                                  | 4 2 2 2 2 3 69 6 3 Wirtschaft erkannt 1 1                   | 1 2 0 1 1 0 3 3 10 69 2 7 Technik erkantt 0 5 1 1         | 0<br>0<br>1<br>20<br>0<br>3<br>3<br>2<br>55<br>Wissenschaft erkant |
| Kultur Gesellschaft Leben Sport Reisen Wirtschaft Technik Wissenschaft Politik Kultur Gesellschaft             | 5 7 2 1 0 5 0 1 Politik erkannt 88 9 17  | 85<br>15<br>5<br>6<br>3<br>2<br>5<br>3<br>4<br>2<br>5<br>3<br>4<br>Wilturerkannt<br>0<br>77 | 2 63 6 6 2 2 2 2 3 0 0 Gesellschaft erkannt 9 4 4 57 3 1          | 3 62 0 0 0 1 1 1 28 Leben erkannt 0 0 1 1 73   | 3 1 3 0 84 2 2 1 1 1 1 Sport erkannt 1 1 3    | 0<br>3<br>6<br>2<br>5<br>84<br>6<br>13<br>7<br>Reisen erkannt<br>0<br>2 | 4 2 2 2 2 3 69 6 3 Wirtschaft erkannt 1 1 2                 | 1 2 0 1 1 0 0 3 1 10 69 2 2 Technik erkannt 0 5 5 1 1 2 2 | 0 0 1 2 20 0 3 3 3 2 2 55 5 Wissenschaft erkant 1 1 4 6 6          |
| Kultur Gesellschaft Leben Sport Reisen Wirtschaft Technik Wissenschaft Politik Kultur Gesellschaft Leben Sport | 5 7 2 1 0 5 0 1 Politik erkannt 88 9 17 4  | a5 15 5 6 3 2 5 3 Kultur erkannt 0 77 9   | 2 63 6 6 2 2 2 2 3 3 0 0 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6          | 3 62 0 0 0 1 1 1 28 Leben erkannt 0 0 1 1 73 0 | 3 1 3 0 04 2 2 1 1 1 5port erkannt 1 3 0 97   | 0 3 6 2 2 5 84 6 13 7 7 Reisen erkannt 0 2 2 6 6 6 2 2 2                | 4 2 2 2 2 3 69 6 3 3 Wirtschaft erkannt 1 1 2               | 1 2 0 1 1 0 0 3 3 10 69 2 Technik erkannt 0 5 1 1 2 0 0   | 0 0 0 1 2 20 0 0 3 3 3 2 2 55 55 Wissenschaft erkann 1 1 4 6 6 0 0 |
| Kultur Gesellschaft Leben Sport Reisen Wirtschaft Technik Wissenschaft Vultur Gesellschaft Leben Sport         | 5 7 2 1 1 0 5 0 1 1 Politik erkannt 88 9 1 1 7 4 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 | 85<br>15<br>5<br>6<br>3<br>2<br>5<br>3<br>Kultur erkannt<br>0<br>77<br>9<br>9               | 2 63 6 6 2 2 2 2 3 1 0 0 Gesellschaft erkannt 9 4 5 5 7 3 0 0 2 2 | 3 62 0 0 1 1 1 28 Leben erkannt 0 0 1 73       | 3 1 3 0 04 2 2 1 1 1 Sport erkannt 1 1 3 0 97 | 0 3 6 2 2 5 84 6 13 7 7 Reisen erkannt 0 2 2 6 6 2 2 8 8 3              | 4 2 2 2 2 2 3 69 6 3 3 Wirtschaft erkannt 1 1 2 2 1 0 0 3 3 | 1 2 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0                   | 0 0 0 1 20 0 0 3 3 3 2 55 5 Wissensthaft erkann 1 1 4 6 6 0 3 3    |

Ohne Wortauswahl72.8%

Worttyp74.4%

Satzglied71.6%

• TF-IDF 75.7%

Preprocessing Methoden der