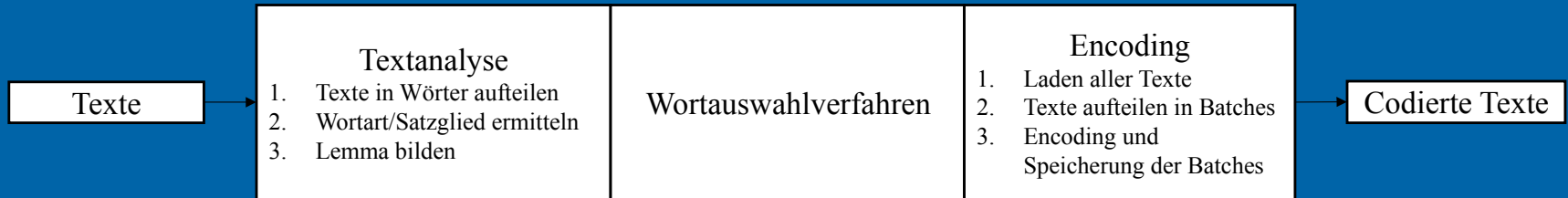


# Verteidigung der Bachelorarbeit

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



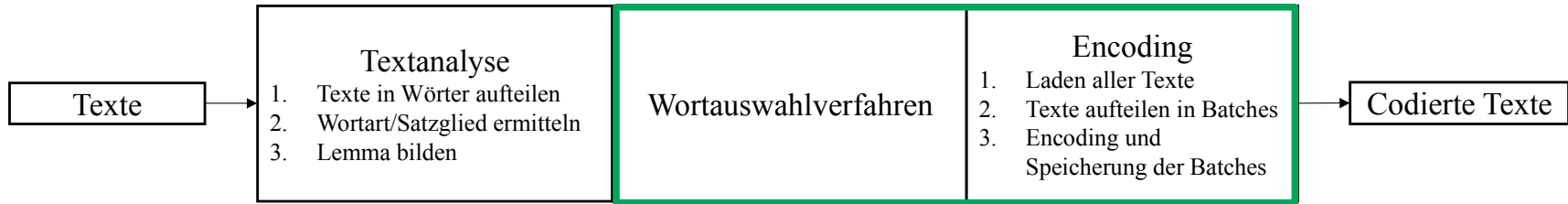
## Beschreibung des Titels

- Textklassifizierung
  - Einsortieren von Texten in zuvor definierte Sparten [1]
  - z.B.: Aufteilung von Zeitungen in Rubriken
- Preprocessing
  - Umwandlung Daten in validen Netinput [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 Spalte
- 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

the	3	3	2
human	2	0	0
build	2	1	0
house	1	5	1
for	4	2	4
shelter	1	0	0
live	1	1	1
in	2	3	2

the	0
human	1
build	2
house	3
for	4
shelter	5
live	6
in	7

3	2	2	3	1	4	1
3	0	1	3	5	2	0
2	0	0	2	1	4	0
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

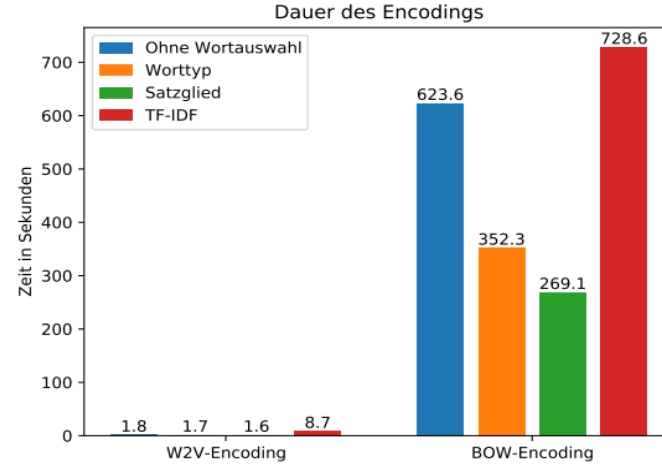
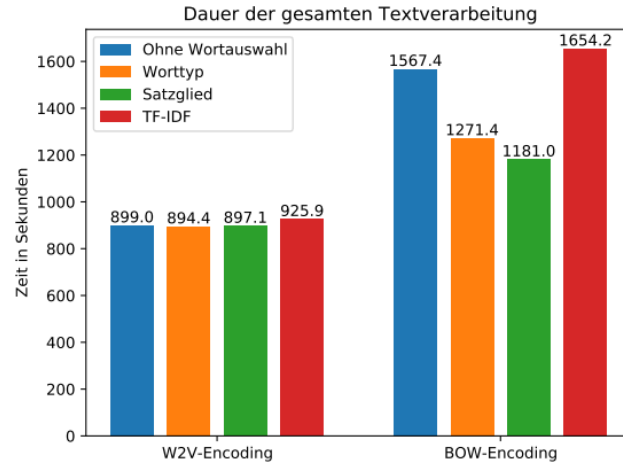
- Trainingsset (9000 Texte), Testset (900 Texte) Preprocessing unterzogen
- Beide Encodings werden in verschiedenen neuronalen Netzen trainiert
- Kombination mit bestem Ergebnis vergleichbare werden gekreuzt verglichen
- Vektorbasiertes Encoding ohne Netz
- Ordinal codierte Texte
- Vergleich Wörterbücher 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	0	19	0	2	0	10	1	1
Kultur	6	77	1	1	0	6	4	4	1
Gesellschaft	9	5	64	1	4	10	2	3	2
Leben	2	4	4	69	0	4	2	2	13
Sport	1	0	1	1	89	5	3	0	0
Reisen	0	1	0	0	0	88	6	5	0
Wirtschaft	3	0	1	1	0	4	81	6	4
Technik	0	1	1	0	0	14	6	77	1
Wissenschaft	0	5	0	26	1	5	4	1	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	1	1	0	1	1	1
Kultur	7	70	11	0	1	3	2	2	4
Gesellschaft	26	4	56	0	2	2	6	1	3
Leben	7	2	11	53	0	0	6	2	19
Sport	2	2	1	3	79	3	6	3	1
Reisen	1	1	4	4	1	65	14	3	7
Wirtschaft	18	1	0	0	0	4	62	9	6
Technik	2	1	1	0	0	3	10	68	15
Wissenschaft	0	0	6	23	2	1	7	2	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	0	0	0	8	7	0	16	0
Kultur	11	4	0	0	36	12	0	37	0
Gesellschaft	17	0	17	0	16	21	0	29	0
Leben	0	0	0	39	8	10	0	43	0
Sport	0	0	0	0	98	0	0	2	0
Reisen	1	0	0	0	5	80	0	14	0
Wirtschaft	7	0	0	0	12	6	15	60	0
Technik	0	0	0	0	1	2	0	97	0
Wissenschaft	0	0	0	9	15	19	0	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	1	4	1	9	0	0
Kultur	6	82	3	0	2	1	3	3	0
Gesellschaft	10	9	57	1	5	10	3	2	3
Leben	1	3	3	58	2	5	2	2	24
Sport	0	6	2	0	73	6	6	6	1
Reisen	0	3	2	6	0	71	5	7	6
Wirtschaft	10	5	2	3	2	6	61	7	4
Technik	2	12	0	1	0	18	10	53	4
Wissenschaft	1	1	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	0	2	4	1	0
Kultur	2	84	0	6	1	3	2	2	0
Gesellschaft	8	9	26	36	2	14	2	0	3
Leben	1	1	0	94	0	1	1	0	2
Sport	1	5	0	4	70	14	4	2	0
Reisen	0	2	0	11	0	75	7	5	0
Wirtschaft	3	3	1	7	0	9	70	5	2
Technik	0	1	0	10	1	33	9	46	0
Wissenschaft	0	1	0	57	0	4	3	0	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	1	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 höhere 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)
- 2) <https://machinelearningknowledge.ai/data-preprocessing-in-machine-learning/>  
(Acessed 1.8.2021)

tu-freiberg.de

 TU Bergakademie Freiberg  bergakademie\_freiberg  TUBergakademie  TUBergakademie

## IMPRESSUM

TU BERGAKADEMIE FREIBERG

Universitätskommunikation

Prüferstr. 2

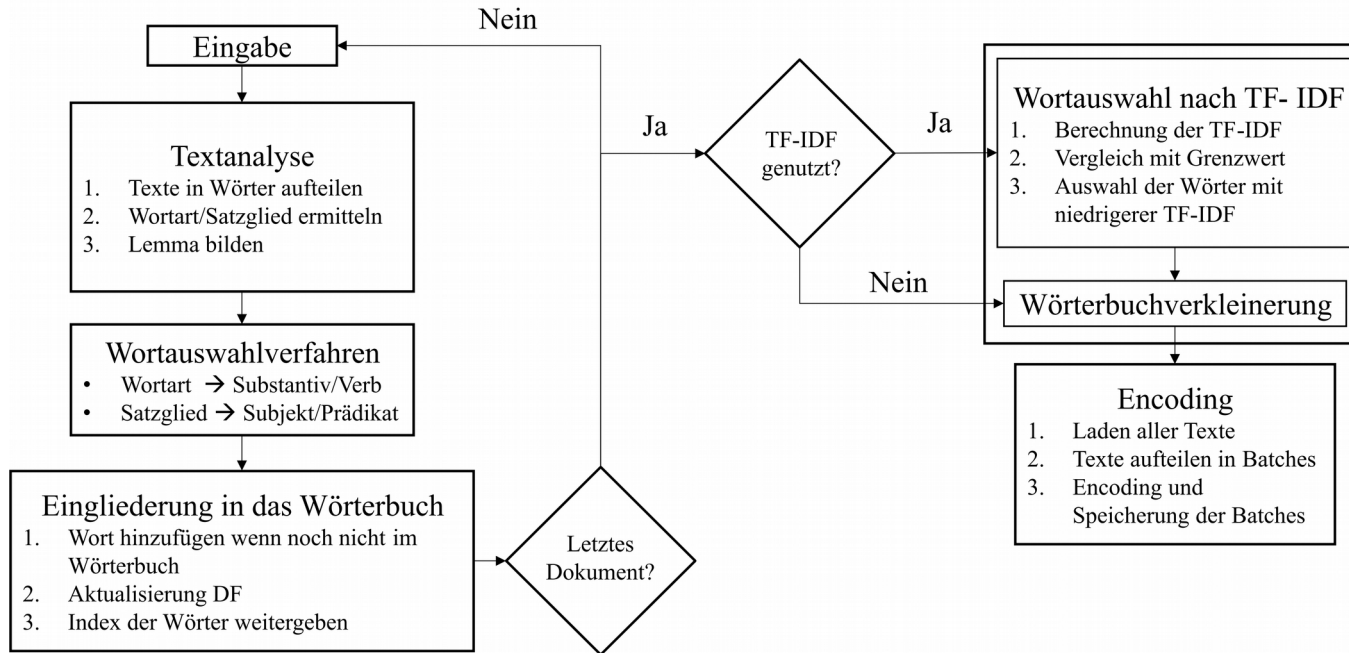
09599 Freiberg

Tel. +49(0)3731 39-2711, -3461

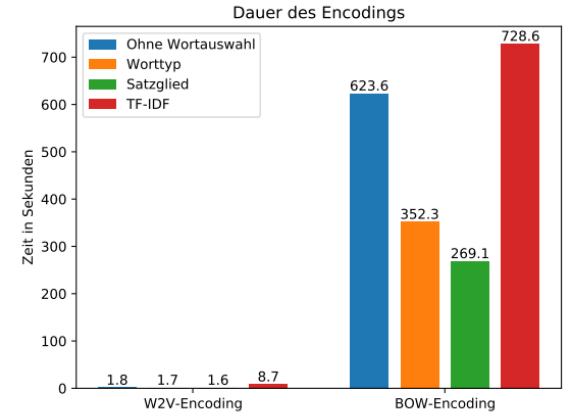
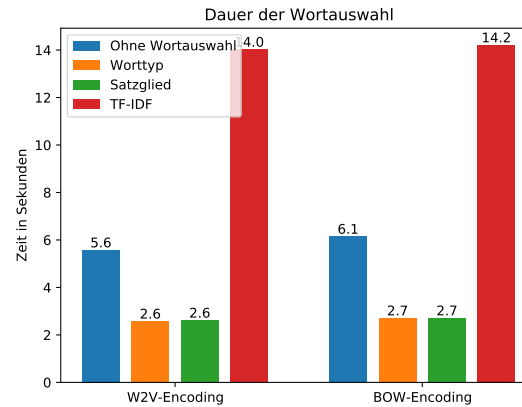
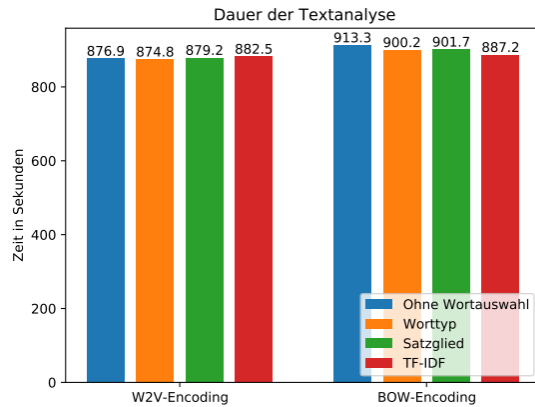
**WELTOFFENE  
HOCHSCHULEN**  
GEGEN FREMDEN-  
FEINDLICHKEIT

 **FAMILIE IN DER  
HOCHSCHULE**

## Aufbau Details



# Zeiten



# Güte

	W2V-Encoding				BOW-Encoding			
	Ohne Wortauswahl	Worttyp	Satzglied	TF-IDF	Ohne Wortauswahl	Worttyp	Satzglied	TF-IDF
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

- Falsch-zuordnungen der Sparte

## Güte

	W2V-Encoding				BOW-Encoding			
	Ohne Wortauswahl	Worttyp	Satzglied	TF-IDF	Ohne Wortauswahl	Worttyp	Satzglied	TF-IDF
Politik	Gesellschaft	Gesellschaft	Gesellschaft	Gesellschaft	Gesellschaft	Gesellschaft	Gesellschaft	Gesellschaft
Kultur	Leben	Gesellschaft	Technik	Reisen	Reisen	Politik	Politik	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

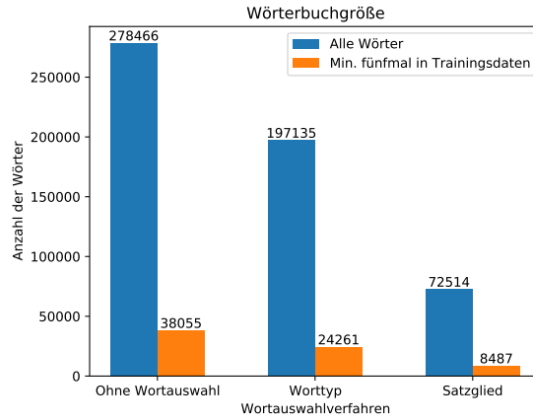
- Spikes  
Falschzuordnungen

	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

# 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

- Falsch-zuordnungen zu Sparte

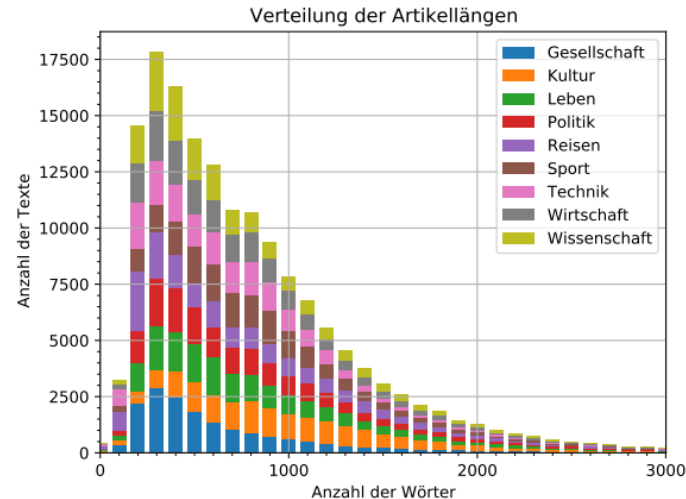


- Wörterbuchgrößen



## 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

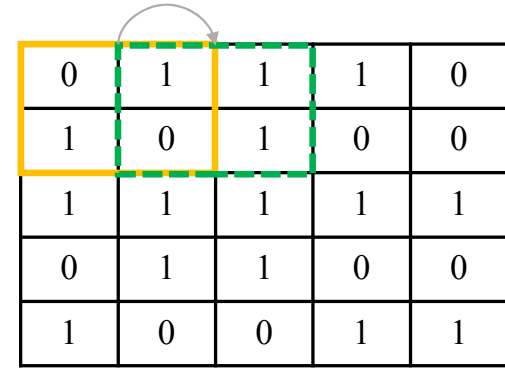
- Termfrequency
  - Vorkommenszahl von Wort in Text
  - Je häufiger Wort in Text → relevanter für Text

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

- Inverse Documentfrequency
  - Anzahl Texte, die Begriff enthalten
  - Je mehr Texte Begriff enthalten → weniger relevant für 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
- **Convolution** verrechnet Matrizen mit Teil der Eingangsdaten verrechnet
- **Pooling** fasst errechnete Daten verlustarm zusammen
- Benötigt Input mit konstanter Dimensionierung



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

## Verwendete Netze

BOW
<b>Convolution Layer</b> <ul style="list-style-type: none"> <li>Filterzahl: 128</li> <li>Fenstergröße: 20</li> <li>Erweiterungsrate: 1</li> <li>Aktivierungsfunktion: Sigmoid</li> <li>Padding: Valid</li> </ul>
<b>MaxPooling Layer</b> <ul style="list-style-type: none"> <li>1D</li> <li>Fenstergröße: 3</li> </ul>
Batchnormalization
Flatten
<b>Dense Layer</b> <ul style="list-style-type: none"> <li>Units: 9</li> <li>Aktivierungsfunktion: Softmax</li> </ul>

W2V
<b>Convolution Layer</b> <ul style="list-style-type: none"> <li>Filterzahl: 64</li> <li>Fenstergröße: 36</li> <li>Erweiterungsrate: 9</li> <li>Aktivierungsfunktion: Relu</li> <li>Padding: Valid</li> </ul>
<b>MaxPooling Layer</b> <ul style="list-style-type: none"> <li>1D</li> <li>Fenstergröße: 2</li> </ul>
Batchnormalization
Flatten
<b>Dense Layer</b> <ul style="list-style-type: none"> <li>Units: 9</li> <li>Aktivierungsfunktion: Softmax</li> </ul>

Ordinal
<b>Convolution Layer</b> <ul style="list-style-type: none"> <li>Filterzahl: 64</li> <li>Fenstergröße: 4</li> <li>Erweiterungsrate: 1</li> <li>Aktivierungsfunktion: Relu</li> <li>Padding: Valid</li> </ul>
<b>MaxPooling Layer</b> <ul style="list-style-type: none"> <li>1D</li> <li>Fenstergröße: 2</li> </ul>
Batchnormalization
Flatten
<b>Dense Layer</b> <ul style="list-style-type: none"> <li>Units: 9</li> <li>Aktivierungsfunktion: Softmax</li> </ul>

## Wortmodifikationen

- Unterschiedlich deklinierte Begriffe zusammenfassen
- Lemmatisierung
  - Ermittlung 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	1	7
Kultur	13	17	11	9	6	15	9	10	10
Gesellschaft	22	13	41	7	3	2	7	1	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	3	8	10	16	18	12	13
Technik	3	7	1	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	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	1	10	3	11	17	18	15	16	9
Wissenschaft	10	18	4	14	13	9	11	12	9

	Politik erkannt	Kultur erkannt	Gesellschaft erkannt	Leben erkannt	Sport erkannt	Reisen erkannt	Wirtschaft erkannt	Technik erkannt	Wissenschaft erkannt
Politik	15	9	20	8	8	14	5	15	6
Kultur	10	23	7	10	10	13	7	12	8
Gesellschaft	21	10	34	8	5	7	5	5	5
Leben	18	10	5	7	12	8	10	16	14
Sport	8	3	6	7	22	11	12	22	9
Reisen	11	11	9	5	8	10	13	15	18
Wirtschaft	7	6	4	3	14	10	18	15	23
Technik	6	6	4	4	24	10	12	20	14
Wissenschaft	9	10	9	8	13	13	11	15	12

	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	1	5
Leben	26	14	6	8	3	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	3	19	9	16	7	14	9

- Ohne Wortauswahl  
18.2%

- Worttyp  
17.3%

- Satzglied  
17.9%

- TF-IDF  
17.2%

## 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	2	0
Gesellschaft	11	8	21	34	4	18	2	0	2
Leben	1	0	0	94	0	1	1	0	3
Sport	0	2	0	3	69	25	1	0	0
Reisen	0	1	0	13	0	78	5	3	0
Wirtschaft	0	0	1	14	0	29	51	3	2
Technik	0	1	0	11	0	50	4	33	1
Wissenschaft	0	1	0	51	0	10	3	0	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	2	2	0
Gesellschaft	8	9	26	36	2	14	2	0	3
Leben	1	1	0	94	0	1	1	0	2
Sport	1	5	0	4	70	14	4	2	0
Reisen	0	2	0	11	0	75	7	5	0
Wirtschaft	3	3	1	7	0	9	70	5	2
Technik	0	1	0	10	1	33	9	46	0
Wissenschaft	0	1	0	57	0	4	3	0	35

	Politik erkannt	Kultur erkannt	Gesellschaft erkannt	Leben erkannt	Sport erkannt	Reisen erkannt	Wirtschaft erkannt	Technik erkannt	Wissenschaft erkannt
Politik	37	4	1	50	0	2	1	3	2
Kultur	2	55	0	23	1	16	1	1	1
Gesellschaft	6	4	30	42	2	15	0	0	1
Leben	1	1	0	79	0	3	0	0	16
Sport	2	4	2	20	45	18	0	5	4
Reisen	0	1	0	23	0	66	2	5	3
Wirtschaft	7	0	1	40	0	12	19	16	5
Technik	2	3	1	17	1	37	2	36	1
Wissenschaft	2	0	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	2	0
Gesellschaft	11	8	21	35	4	17	2	0	2
Leben	1	0	0	94	0	1	1	0	3
Sport	0	2	0	4	69	24	1	0	0
Reisen	0	1	0	12	0	79	5	3	0
Wirtschaft	0	0	1	13	0	28	53	3	2
Technik	0	1	0	10	0	45	4	39	1
Wissenschaft	0	1	0	50	0	10	4	0	35

- Ohne Wortauswahl  
56.3%
- Worttyp  
61.1%
- Satzglied  
45.7%
- TF-IDF  
57.7%



## Vektorrepräsentation

	Politik erkannt	Kultur erkannt	Gesellschaft erkannt	Leben erkannt	Sport erkannt	Reisen erkannt	Wirtschaft erkannt	Technik erkannt	Wissenschaft erkannt
Politik	55	12	20	2	1	0	2	1	7
Kultur	4	85	2	5	0	0	3	0	1
Gesellschaft	5	9	66	5	1	4	3	0	7
Leben	2	3	4	71	0	0	3	0	17
Sport	1	3	2	1	90	1	0	0	2
Reisen	1	7	5	6	4	48	6	5	18
Wirtschaft	6	1	3	5	5	3	36	3	38
Technik	1	6	1	6	0	5	5	60	16
Wissenschaft	0	3	1	31	3	0	5	0	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	0	1	3	2	2	4
Gesellschaft	26	4	56	0	2	2	6	1	3
Leben	7	2	11	53	0	0	6	2	19
Sport	2	2	1	3	79	3	6	3	1
Reisen	1	1	4	4	1	65	14	3	7
Wirtschaft	18	1	0	0	0	4	62	9	6
Technik	2	1	1	0	0	3	10	68	15
Wissenschaft	0	0	6	23	3	1	7	2	59

	Politik erkannt	Kultur erkannt	Gesellschaft erkannt	Leben erkannt	Sport erkannt	Reisen erkannt	Wirtschaft erkannt	Technik erkannt	Wissenschaft erkannt
Politik	24	4	31	0	4	0	10	24	3
Kultur	1	54	1	0	17	4	3	20	0
Gesellschaft	4	4	56	1	12	8	3	6	6
Leben	1	1	8	36	0	3	1	13	37
Sport	0	1	1	0	86	2	1	6	3
Reisen	0	3	3	0	3	67	9	8	7
Wirtschaft	2	1	4	1	1	6	63	19	3
Technik	0	0	1	0	2	5	3	88	1
Wissenschaft	0	0	2	8	2	4	2	9	73

	Politik erkannt	Kultur erkannt	Gesellschaft erkannt	Leben erkannt	Sport erkannt	Reisen erkannt	Wirtschaft erkannt	Technik erkannt	Wissenschaft erkannt
Politik	59	10	17	1	2	1	6	3	1
Kultur	6	72	3	0	1	8	4	4	2
Gesellschaft	17	10	55	3	3	4	2	2	4
Leben	3	1	7	64	0	2	4	2	17
Sport	0	3	3	1	79	10	2	2	0
Reisen	3	3	0	2	0	67	11	10	4
Wirtschaft	4	0	1	7	0	21	48	11	8
Technik	1	3	2	6	1	11	10	59	7
Wissenschaft	0	0	1	26	0	5	4	2	62

- Ohne Wortauswahl  
63.1%

- Worttyp  
66.2%

- 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	1	7	2	5	1
Gesellschaft	6	8	63	2	3	12	3	1	2
Leben	0	3	2	83	0	5	1	0	6
Sport	0	1	0	1	84	12	2	0	0
Reisen	1	0	0	0	0	90	4	4	1
Wirtschaft	2	0	1	2	0	10	76	5	4
Technik	0	1	1	3	0	26	4	64	1
Wissenschaft	0	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 erkannt
Politik	67	0	19	0	2	0	10	1	1
Kultur	6	77	1	1	0	6	4	4	1
Gesellschaft	9	5	64	1	4	10	2	3	2
Leben	2	4	4	69	0	4	2	2	13
Sport	1	0	1	1	89	5	3	0	0
Reisen	0	1	0	0	0	88	6	5	0
Wirtschaft	3	0	1	1	0	4	81	6	4
Technik	0	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 erkannt
Politik	73	5	13	1	3	0	4	1	0
Kultur	5	85	2	0	1	3	2	2	0
Gesellschaft	7	15	63	3	3	6	2	0	1
Leben	2	5	6	62	0	2	2	1	20
Sport	1	6	2	0	84	5	2	0	0
Reisen	0	3	2	0	2	84	3	3	3
Wirtschaft	5	2	2	1	2	6	69	10	3
Technik	0	5	3	1	1	13	6	69	2
Wissenschaft	1	3	0	28	1	7	3	2	55

	Politik erkannt	Kultur erkannt	Gesellschaft erkannt	Leben erkannt	Sport erkannt	Reisen erkannt	Wirtschaft erkannt	Technik erkannt	Wissenschaft erkannt
Politik	88	0	9	0	1	0	1	0	1
Kultur	9	77	4	0	1	2	1	5	1
Gesellschaft	17	9	57	1	3	6	2	1	4
Leben	4	9	3	73	0	2	1	2	6
Sport	0	1	0	0	97	2	0	0	0
Reisen	1	2	2	0	0	83	3	6	3
Wirtschaft	5	1	1	2	1	3	69	12	6
Technik	1	3	2	1	0	5	0	85	3
Wissenschaft	0	9	0	28	2	6	2	1	52

- Ohne Wortauswahl  
72.8%

- Worttyp  
74.4%

- Satzglied  
71.6%

- TF-IDF  
75.7%