Entscheidungsbäume

Masterseminar Klassifikation und Clustering

Tim Sockel, Tatiana Enina

Dozent: PD Dr. Stefan Langer

12.12.2022

Gliederung

1. Definition
2. Der Algorithmus
3. Vor- und Nachteile des Algorithmus
4. Implementierung
5. Literatur

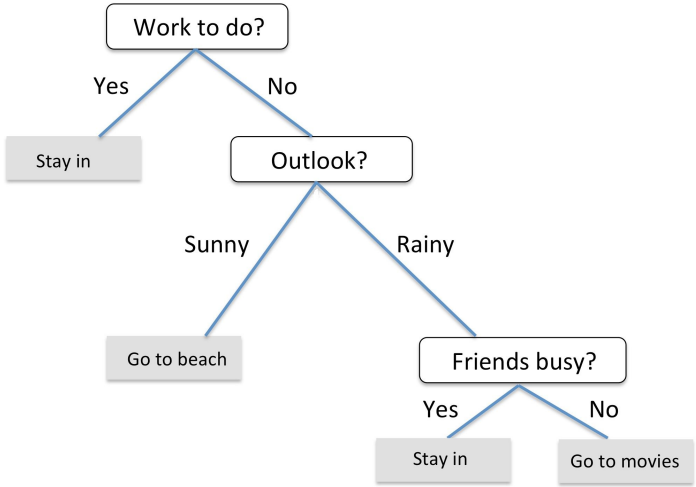
Definition

Entscheidungsbaum – ein Lernalgorithmus, bei dem Klassifikations- oder Regressionsmodelle in Form einer Baumstruktur aufgebaut werden.

⇒ Ein Ansatz zur *mehrstufigen Entscheidungsfindung*: eine komplexe Entscheidung wird in eine Vereinigung mehrerer einfacherer Entscheidungen zerlegt.

Geschichte:

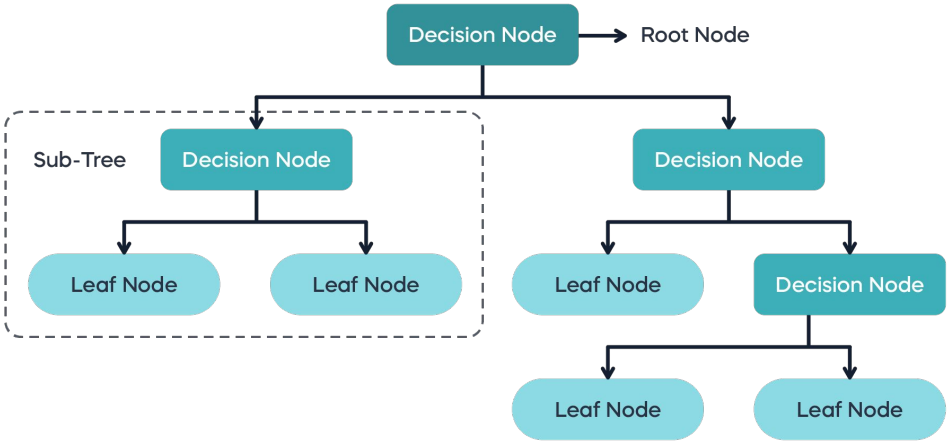
- 1963 – AID project (Morgan, Sonquist): erster Algorithmus mit *Regression*
- 1972 – THAID project (Messenger, Mandell): erster Algorithmus zur *Klassifikation*



Ein Entscheidungsbaum: Beispiel (https://tjmachinelearning.com/lectures/1718/dct/, Zugriff: 12.12.2022)

Definition

- einfache hierarchische, baumartige Struktur;
- **Bestandteile**:
  - ein Wurzelknoten (eng. *root node*),
  - endliche Menge von internen Knoten (eng. *internal* oder *decision nodes*),
  - endliche Menge von Blattknoten (eng. *leafs* oder *terminal nodes*).
- Prozess der Konstruierung eines Baums – Bauminduktion, Baumaufbau (eng. *induction, tree building* oder ***tree growing***);
- meistens wird die **Top-Down**-Methode verwendet.



Struktur eines Entscheidungsbaums (https://365datascience.com/tutorials/machine-learning-tutorials/decision-trees/, Zugriff: 10.12.2022)

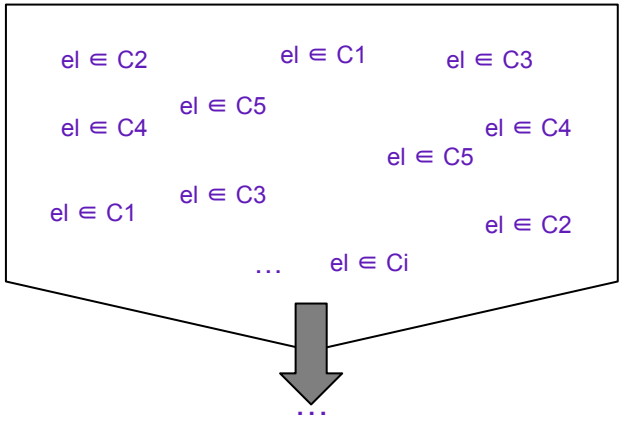
Der Algorithmus

Trainingsdaten:

- Menge *S* aus *n* Elementen; jedes Element gehört zu einer Klasse (C_1, …, C_i)

- Attribute: A_1, …, A_j

- Werte: (a_1,…, a_p)



Der Algorithmus

Trainingsdaten:

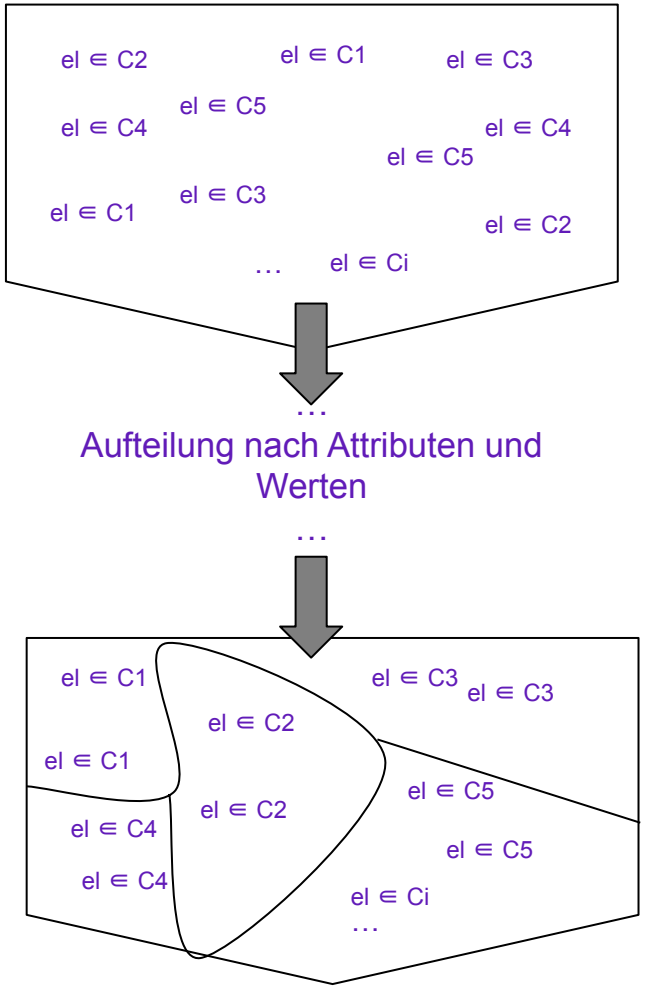
- Menge *S* aus *n* Elementen; jedes Element gehört zu einer Klasse $(C_1, \ldots, C_i)$

- Attribute: $A_1, \ldots, A_j$

- Werte: $(a_1, \ldots, a_p)$

Ziel:

Aufgeteilte Menge *S*, bei der Elemente jeder Teilmenge eine (möglichst) reine Klasse repräsentieren



Der Algorithmus

Probleme:

- Generieren (eng. *growing*): Auswahl der Attribute
- Auswahl des Lernstoppkriteriums: Stutzen (eng. *pruning*)

```
TreeGrowing (S,A,y)
Where:
S - Training Set
A - Input Feature Set
y - Target Feature
Create a new tree T with a single root node.
IF One of the Stopping Criteria is fulfilled THEN
    Mark the root node in T as a leaf with the most
    common value of y in S as a label.
ELSE
    Find a discrete function f(A) of the input
        attributes values such that splitting S
        according to f(A)'s outcomes (v1,...,vn) gains
        the best splitting metric.
    IF best splitting metric > treshold THEN
        Label t with f(A)
        FOR each outcome vi of f(A):
            Set Subtreei= TreeGrowing (σf(A)=viS,A,y).
            Connect the root node of tT to Subtreei with
                an edge that is labelled as vi
        END FOR
    ELSE
        Mark the root node in T as a leaf with the most
            common value of y in S as a label.
    END IF
END IF
RETURN T
TreePruning (S,T,y)
Where:
S - Training Set
y - Target Feature
T - The tree to be pruned
DO
    Select a node t in T such that pruning it
        maximally improve some evaluation criteria
    IF t≠∅ THEN T=pruned(T,t)
UNTIL t=∅
RETURN T
```

Algorithmischer Top-Down Framework für die Induktion von Entscheidungsbäumen (Rokach, Lior & Maimon, 2005: 169)

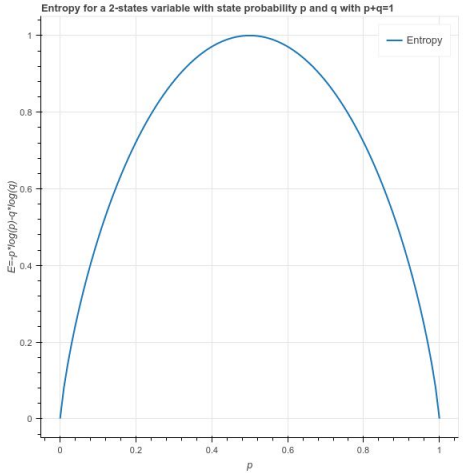
Der Algorithmus: Generierung eines Baums

Entropie:

$$E(S) = -\sum_i p_i \log p_i = -\sum_{i=1}^{n} \frac{N_i}{N} \log\left(\frac{N_i}{N}\right)$$

n – die Anzahl der Klassen, N_i – die Anzahl der Elemente einer Klasse *i, N* – die Gesamtzahl der Elemente.

E(*S*)= 0, falls p_i = 1 für eine Klasse *i* in *S*
E(*S*)= 1 für 2 Klassen, falls p_1, p_2 = 1/2



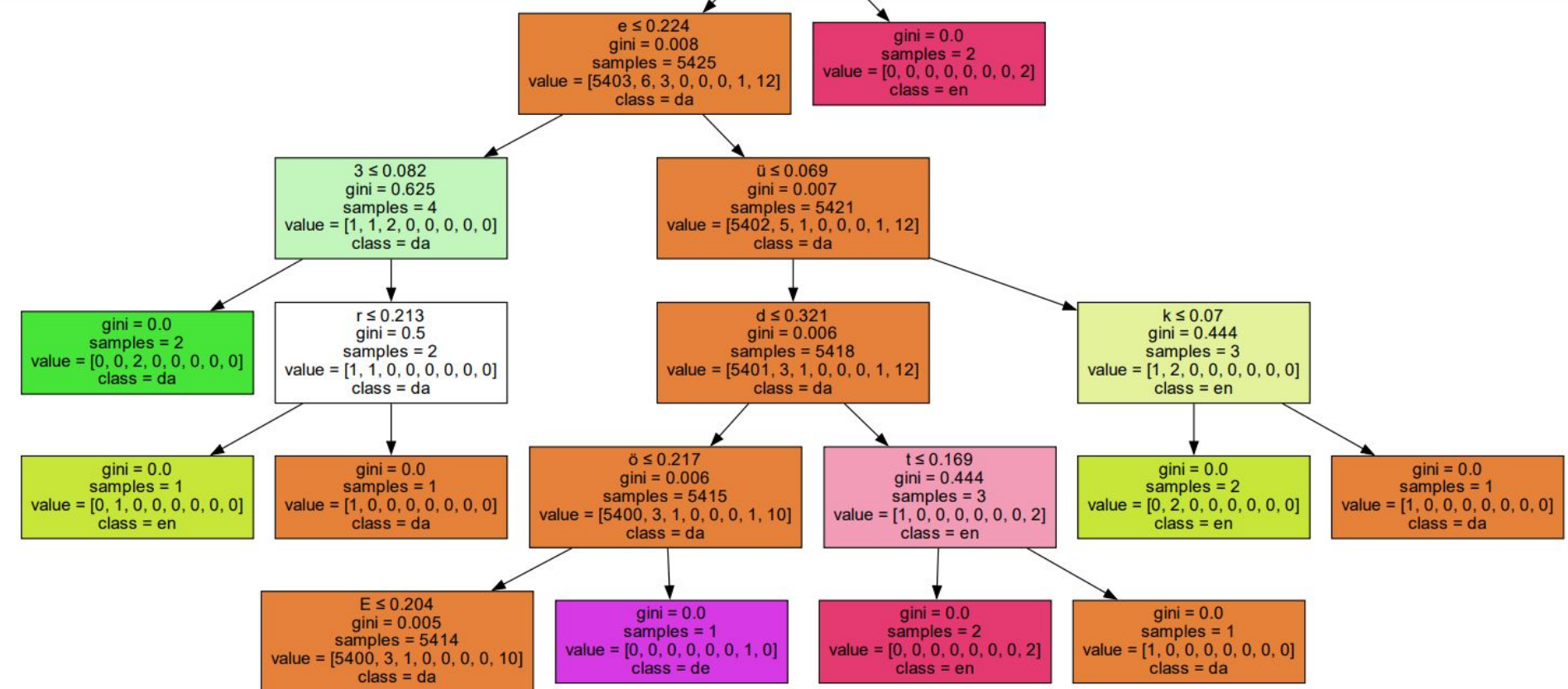
Grafik für $E = p*\log(p)-q*\log(q)$, $q+p=1$ (https://bricaud.github.io/personal-blog/entropy-in-decision-trees/, Zugriff: 8.12.22)

Gini-Index:

$$\text{Gini}(S) = 1 - \sum_{i=1}^{n} p_i^2$$

n - die Anzahl der Klassen,
p_i - die Wahrscheinlichkeit einer Klasse i

kleiner Gini-Index ⇔ geringe Unreinheit
großer Gini-Index ⇔ hohe Unreinheit



Visualisierung eines Entscheidungsbaums zur Kategorisierung von Briefen nach Sprache (erstellt mit *Graphviz*)

Der Algorithmus: Beispiele von Klassifikatoren

- ID3: Entropie, Informationsgewinn (eng. *Information gain*);
- C4.5: Informationsgewinn;
- CART (Classification and Regression Tree): Gini Index.

Der Algorithmus: Abbruch des Lernprozesses

Pre-Pruning:

- Baumtiefenbegrenzung (`max_depth`)
- Beste Mindestanzahl der Elemente bei internen (Split-)Knoten (`min_samples_split`)
- Beste Mindestanzahl der Elemente bei Endknoten (`min_samples_leaf`)

Post-Pruning (z. B. ‘Minimal Cost-Complexity Pruning’):

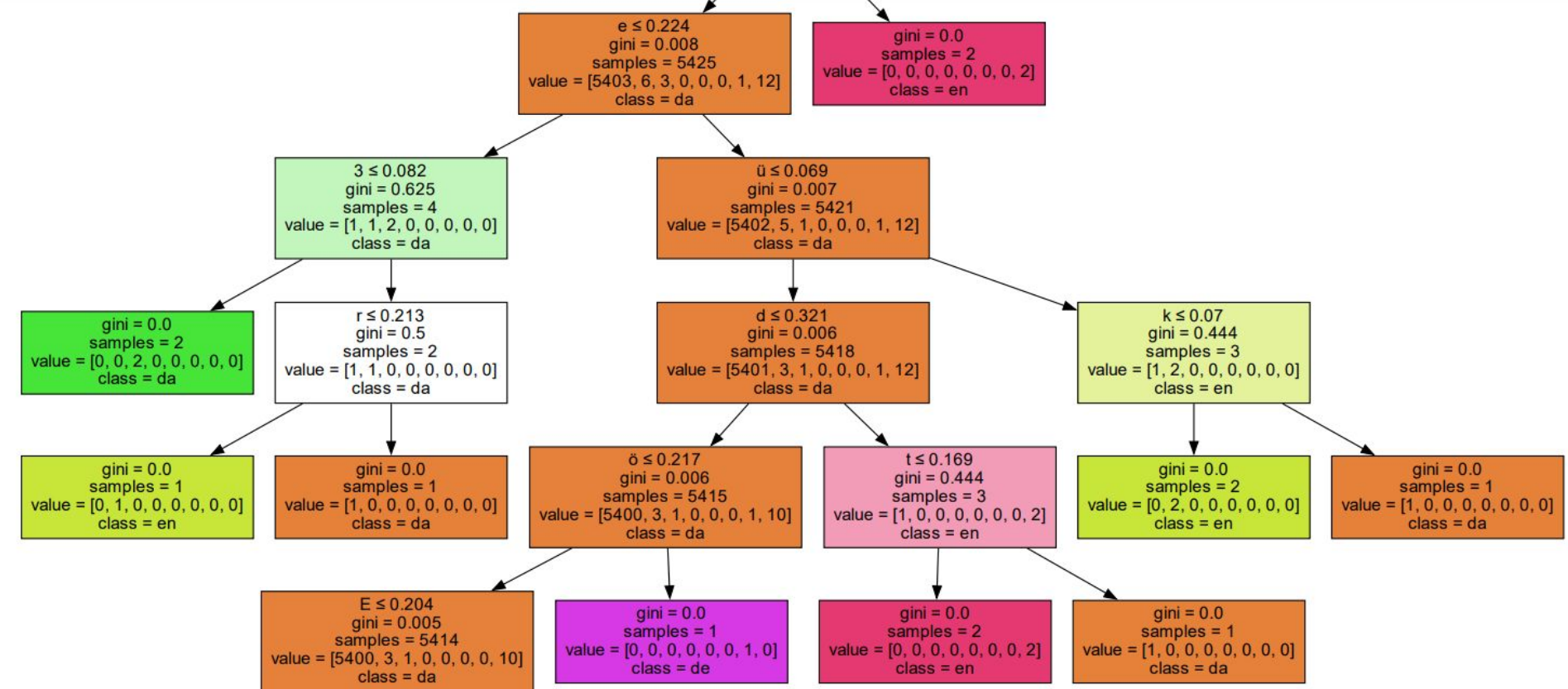
- Entscheidungsbaum wird zunächst mit voller Baumtiefe generiert
- `ccp_alpha`-Wert kann darauf basierend bspw. mittels einer Funktion berechnet oder geschätzt werden
- Entscheidungsbaum kann mittels verschiedenen `ccp_alpha`-Werten trainiert werden und die Evaluationsergebnisse können verglichen werden

Vorteile

- Weniger Datenvorbereitung während der Vorverarbeitung;
- Merkmalsskalierung nicht erforderlich;
- schneller Trainingsprozess;
- können automatisch mit fehlenden Werten umgehen;
- leicht zu verstehen und visualisieren.

Nachteile

- Hinzufügen neuer Datenpunkte führt zur Neugenerierung des Gesamtbaums (Knoten müssen neu berechnet werden);
- werden durch Noise beeinflusst und instabil;
- das Ergebnis ist teils von der Reihenfolge der eingegebenen Texte abhängig;
- für große Datasets nicht geeignet, Gefahr von Overfitting.
  → Lösung: Stutzen (*pruning*).



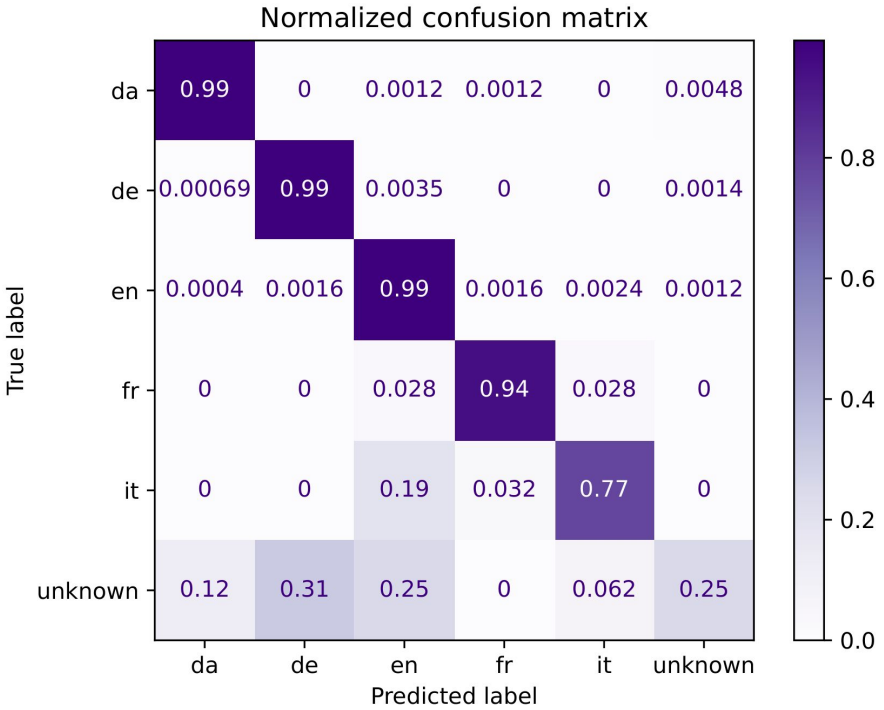
Visualisierung eines Entscheidungsbaums zur Kategorisierung von Briefen nach Sprache (erstellt mit *Graphviz*)

Implementierung

Briefsammlung – Kategorisierung nach Sprache

Methode: Zeichenbasierte TF-IDF-Matrix; ohne sog. *Lowercasing*

	Precision	Recall	F1	Support
da	1.00	0.99	0.99	838
de	0.99	0.99	0.99	1443
en	0.99	0.99	0.99	2517
fr	0.85	0.94	0.89	36
it	0.75	0.77	0.76	31
unk.	0.31	0.25	0.28	16
acc.			0.99	4881
macro	0.81	0.82	0.82	4881
weight.	0.99	0.99	0.99	4881

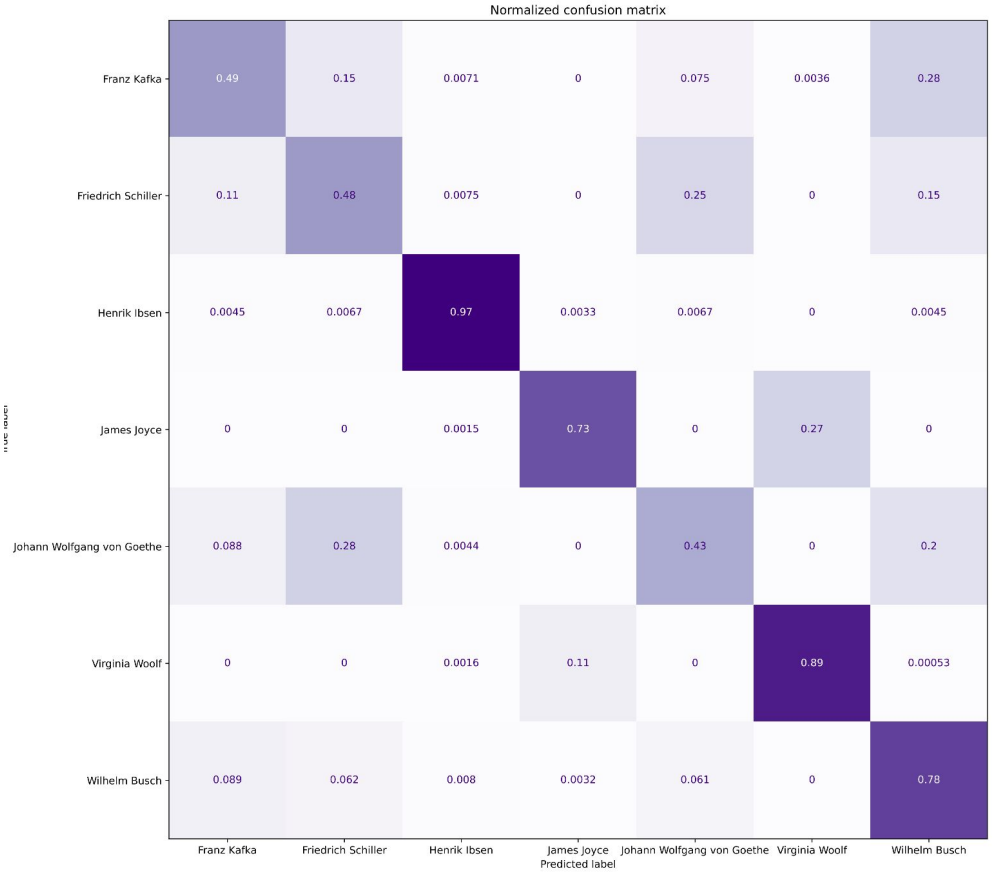


Implementierung

Briefsammlung – Kategorisierung nach Autorinnen und Autoren

Methode: Wortbasierte TF-IDF-Matrix

	Prec.	Recall	F1	Support
Kafka	0.56	0.49	0.52	280
Schiller	0.46	0.48	0.47	266
Ibsen	0.98	0.97	0.98	897
Joyce	0.69	0.73	0.71	682
Goethe	0.43	0.43	0.43	228
Woolf	0.90	0.89	0.89	1901
Busch	0.75	0.78	0.76	627
acc.			0.80	4881
macro	0.68	0.68	0.68	4881
weight.	0.80	0.80	0.80	4881

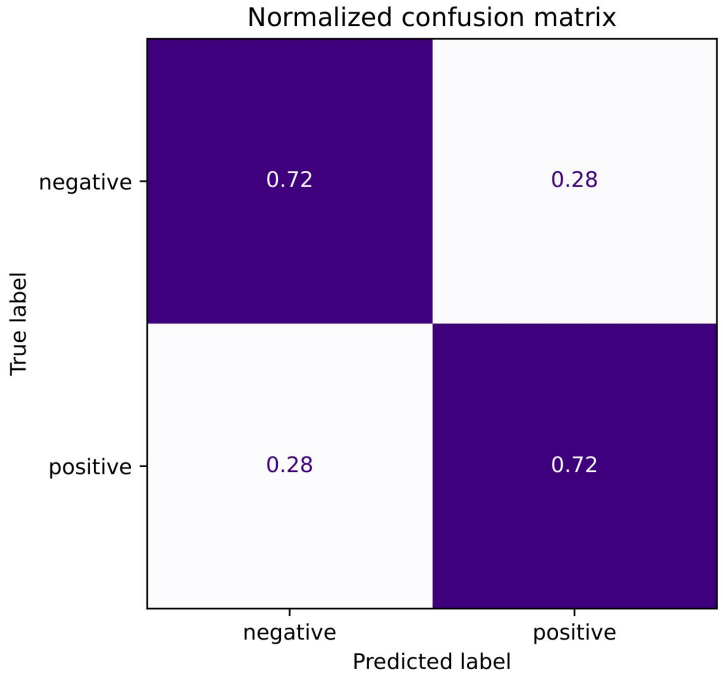


Implementierung

IMDb-Filmrezensionen – Sentiment-Analyse

Methode: TF-IDF-Matrix aus Uni- und Bigrammen;
45.000 häufigste Merkmale

	Precision	Recall	F1	Support
Negativ	0.72	0.72	0.72	4985
Positiv	0.72	0.72	0.72	5015
acc.			0.72	10000
macro	0.72	0.72	0.72	10000
weight.	0.72	0.72	0.72	10000



Implementierung

Huffington Post – Kategorisierung nach Ressorts

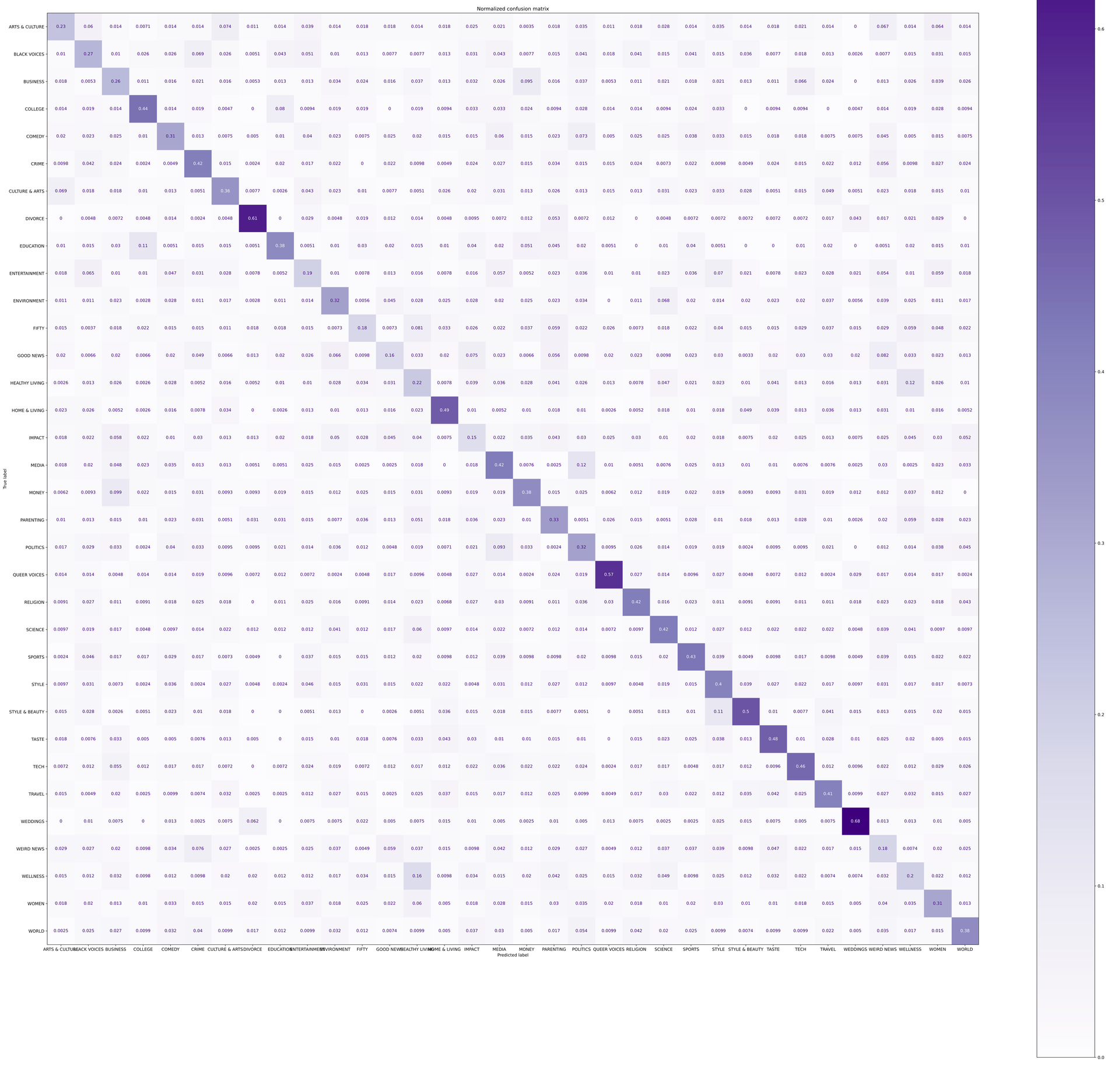
Methode: Wortbasierte TF-IDF-Matrix; ohne englische Stoppwörter

➔ weitere Optimierung mittels *Minimal Cost-Complexity Pruning* (`ccp_alpha` = 0.00006)

⇒ Accuracy von 36 % konnte auf 41 % verbessert werden

	Prec.	Recall	F1	Support
Arts & Culture	0.27	0.23	**0.25**	282
Black Voices	0.29	0.27	**0.28**	392
Business	0.26	0.26	**0.26**	380
College	0.40	0.44	**0.42**	212
Comedy	0.32	0.31	**0.32**	399
Crime	0.40	0.42	**0.41**	409
Culture & Arts	0.40	0.36	**0.38**	391
Divorce	0.68	0.61	**0.64**	419
Education	0.33	0.38	**0.36**	198
Entertainment	0.22	0.19	**0.21**	387
Environment	0.31	0.32	**0.32**	355
Fifty	0.20	0.18	**0.19**	273
Good News	0.19	0.16	**0.17**	305
Healthy Living	0.19	0.22	**0.20**	386
Home & Living	0.51	0.49	**0.50**	386
Impact	0.17	0.15	**0.16**	400
Media	0.32	0.42	**0.36**	395
Money	0.36	0.38	**0.37**	324
Parenting	0.30	0.33	**0.32**	391
Politics	0.29	0.32	**0.30**	420
Queer Voices	0.63	0.57	**0.60**	415
Religion	0.48	0.42	**0.45**	440
Science	0.41	0.42	**0.42**	414
Sports	0.41	0.43	**0.42**	410
Style	0.34	0.40	**0.37**	413
Style & Beauty	0.52	0.50	**0.51**	391
Taste	0.47	0.48	**0.48**	397
Tech	0.46	0.46	**0.46**	416
Travel	0.41	0.41	**0.41**	405
Weddings	0.68	0.68	**0.68**	400
Weird News	0.17	0.18	**0.18**	408
Wellness	0.21	0.20	**0.20**	407
Women	0.29	0.31	**0.30**	400
World	0.40	0.38	**0.39**	404
accuracy			**0.36**	12824
macro avg	0.36	0.36	**0.36**	12824
weighted avg	0.37	0.36	**0.37**	12824

Modellqualität ohne Pruning


Normalized confusion matrix
True label
Predicted label
ARTS & CULTURE
BLACK VOICES
BUSINESS
COLLEGE
COMEDY
CRIME
CULTURE & ARTS
DIVORCE
EDUCATION
ENTERTAINMENT
ENVIRONMENT
FIFTY
GOOD NEWS
HEALTHY LIVING
HOME & LIVING
IMPACT
MEDIA
MONEY
PARENTING
POLITICS
QUEER VOICES
RELIGION
SCIENCE
SPORTS
STYLE
STYLE & BEAUTY
TASTE
TECH
TRAVEL
WEDDINGS
WEIRD NEWS
WELLNESS
WOMEN
WORLD
0.6
0.5
0.4
0.3
0.2
0.1
0.0

Implementierung

Huffington Post – Kategorisierung nach Ressorts

Methode: Wortbasierte TF-IDF-Matrix; ohne englische Stoppwörter

➔ weitere Optimierung mittels *Minimal Cost-Complexity Pruning* (`ccp_alpha` = 0.00006)

⇒ Accuracy von 36 % konnte auf 41 % verbessert werden

	Prec.	Recall	F1	Support
Arts & Culture	0.52	0.25	**0.34**	282
Black Voices	0.53	0.31	**0.39**	392
Business	0.34	0.32	**0.33**	380
College	0.51	0.53	**0.52**	212
Comedy	0.54	0.28	**0.37**	399
Crime	0.50	0.47	**0.48**	409
Culture & Arts	0.52	0.37	**0.43**	391
Divorce	0.77	0.63	**0.69**	419
Education	0.43	0.44	**0.44**	198
Entertainment	0.10	0.42	**0.16**	387
Environment	0.44	0.37	**0.40**	355
Fifty	0.19	0.17	**0.18**	273
Good News	0.38	0.11	**0.18**	305
Healthy Living	0.20	0.19	**0.20**	386
Home & Living	0.38	0.60	**0.47**	386
Impact	0.18	0.16	**0.17**	400
Media	0.51	0.43	**0.47**	395
Money	0.45	0.47	**0.46**	324
Parenting	0.34	0.54	**0.42**	391
Politics	0.34	0.43	**0.38**	420
Queer Voices	0.80	0.61	**0.69**	415
Religion	0.63	0.45	**0.52**	440
Science	0.49	0.47	**0.48**	414
Sports	0.49	0.48	**0.48**	410
Style	0.43	0.38	**0.40**	413
Style & Beauty	0.53	0.49	**0.51**	391
Taste	0.60	0.52	**0.56**	397
Tech	0.54	0.47	**0.50**	416
Travel	0.45	0.42	**0.44**	405
Weddings	0.77	0.72	**0.74**	400
Weird News	0.24	0.17	**0.20**	408
Wellness	0.24	0.27	**0.25**	407
Women	0.37	0.37	**0.37**	400
World	0.52	0.36	**0.43**	404
accuracy			**0.41**	12824
macro avg	0.45	0.40	**0.41**	12824
weighted avg	0.45	0.41	**0.42**	12824

Modellqualität mit Pruning

Normalized confusion matrix

True label

Predicted label

ARTS & CULTURE
BLACK VOICES
BUSINESS
COLLEGE
COMEDY
CRIME
CULTURE & ARTS
DIVORCE
EDUCATION
ENTERTAINMENT
ENVIRONMENT
FIFTY
GOOD NEWS
HEALTHY LIVING
HOME & LIVING
IMPACT
MEDIA
MONEY
PARENTING
POLITICS
QUEER VOICES
RELIGION
SCIENCE
SPORTS
STYLE
STYLE & BEAUTY
TASTE
TECH
TRAVEL
WEDDINGS
WEIRD NEWS
WELLNESS
WOMEN
WORLD

0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0

Literatur

- Loh, W. 2014. Fifty Years of Classification and Regression Trees. *International Statistical Review*. International Statistical Institute (ISI). 82, 3, 329–348.
- Manning, C. and Schütze, H. 2001. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge MA.
- Murthy, S. 1997. Automatic construction of decision trees from data: A Multi-disciplinary survey. Kluwer Academic Publishers, Boston.
- Rokach, L. and Maimon, O. 2005. Decision Trees. *The Data Mining and Knowledge Discovery Handbook*, 165-192. Springer, Boston, MA.
- Pedregosa et al. 2011. Scikit-learn: Machine Learning in Python. *JMLR 12*. 2825-2830.
- Ellson J. et al. 2004. *Graphviz and Dynagraph — Static and Dynamic Graph Drawing Tools.*