Klassifikation der Texte mit K-Nearest-Neighbor (KNN) Algorithmus

Ludwig Maximilians Universität München (LMU)
Centrum für Infrmations- und Sprachverarbeitung (CIS)
Masterseminar: Klassifikation und Clustering
Dozent: Stefan Langer
Referat von: Galyna Gerasymchuk
WiSe 23/24

22. Januar 2024

Inhalt

Wichtige Daten:

1951 - NICHTPARAMETRISCHE MUSTERKLASSIFIKATION

Evelyn Fix und Joseph Lawson Hodges, Jr. Die nichtparametrische Musterklassifikation, aus der später kNN werden sollte, wird erstmals in einem unveröffentlichten technischen Bericht der US Air Force vorgestellt - wahrscheinlich aus Gründen des Datenschutzes nach dem Zweiten Weltkrieg - wurde nie offiziell veröffentlicht.

1967 - NEAREST NEIGHBOUR PATTERN KLASSIFIKATION

Thomas Cover und Peter Hart veröffentlichen "Nearest Neighbour Pattern Classification", in dem sie eine Obergrenze für die Fehlerrate der Mehrklassenklassifikation kNN nachweisen.

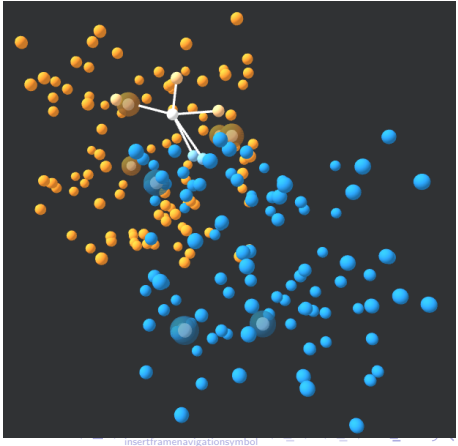
1985 - FUZZY KNN

James Keller entwickelte eine 'unscharfe' Version des kNN-Algorithmus. Aufgrund seiner niedrigeren Fehlerquote ist er mit komplexeren Mustererkennungsverfahren vergleichbar.

KNN - Einführung

K-Nearest Neighbors-Algorithmus (KNN)

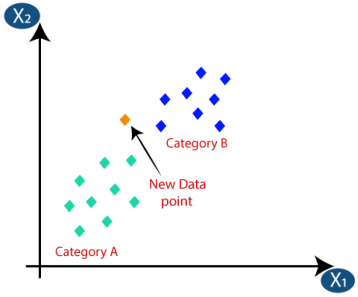
- 'nichtlinearer' Klassifikationsalgorithmus
- KNN basiert auf der Idee, dass ähnliche Datenpunkte im Merkmalsraum tendenziell denselben Klassen zugeordnet werden
- überwachter ML-Algorithmus - für Klassifizierungs- und für Regressionsvorhersage
- nicht-parametrischer und 'fauler' Algorithmus

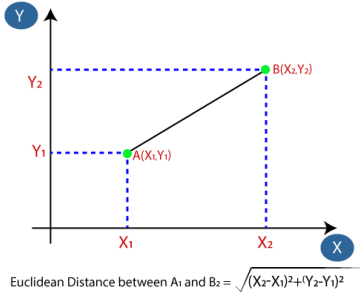


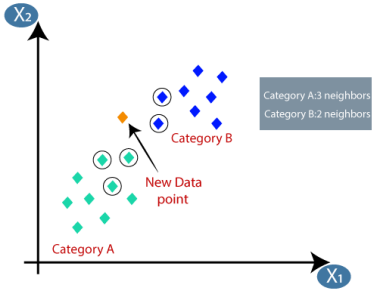
KNN - Implementierungsschritte:

1. Vorbereitung des Trainingsdatensatzes, der aus markierten (labeled) Instanzen (Datenpunkten) mit bekannten Klassenbezeichnungen besteht
2. Bestimmung der Anzahl der k-Nachbarn **(k)**: $k \geq 1$
   Es soll eine **ungerade** Zahl sein!
3. Auswahl einer **Distanzmetrik**
   - **Euclidean** Distanz $= \sqrt{(p1_x - p2_x)^2 + (p1_y - p2_y)^2}$
   - Manhattan Distanz (L1-Norm)
   - Minkowski Distanz
4. Berechnung der Distanz
5. Cross-Validation
   - k-Fold Cross-Validation, z.B. $num\_folds = 5$
6. Optimierung den Hyperparameter **k**
7. Validierung der Ergebnissen

KNN - wie es funktioniert:







Am Ende wird die vorhergesagte Klassenbezeichnung durch die Abstimmung für die nächsten Nachbarn bestimmt und zurückgegeben

Anwendungsgebiete von KNN

- Mustererkennung
- Klassifikation und Regression
- Text- und Dokumentklassifikation
- Empfehlungssysteme
- Handschriftenerkennung
- Bilderkennung
- Gesundheitswesen - diagnostizieren von Krankheiten
- ...

Vor- und Nachteile des k-Nearest Neighbor Algorithmus

Vorteile:

- einfach zu verstehen und zu implementieren;
- kein Training notwendig;
- ein Hyperparameter k;
- adaptivität an die Daten;
- mehrere Möglichkeiten zum berechnen

Nachteile:

- rechenintensiv und langsam, je größer der Datensatz ist;
- hoher Speicherbedarf;
- großes Problem: welches k ist sinnvoll?
- empfindlich gegenüber irrelevanten Merkmalen und Ausreißern;
- nicht lernfähig

Imlementierung & Evaluierung

Implementierungsschritte:

1. **load_data**
2. TF_IDF Vektorisierung
   - **TfidfVectorizer** - es dient dazu, Textdaten in numerische Vektoren umzuwandeln
3. Training: **KNeighborsClassifier**
   - *n_neighbors=k_neighbors* (1, n, 257)
   - *metric='euclidean'*
4. Evaluierung und Ergebnisse:
   - Recall
   - Precision
   - F1
   - Accuracy
   - Micro-Average
   - Macro-Average
   - Konfusionsmatrix

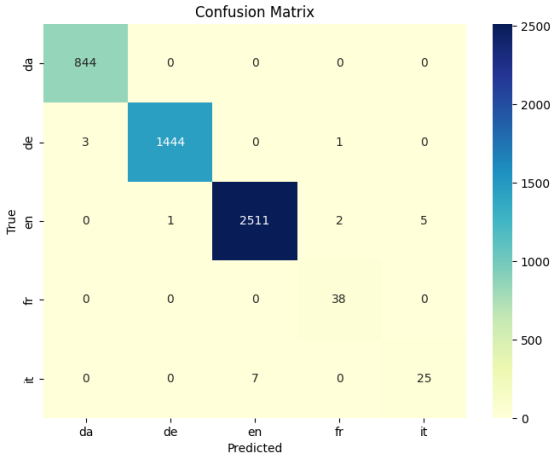
Imlementierung & Evaluierung: Sprachklassifizierung

Klassifikation nach Sprache mit **k=1**

Classification Report:

- **Accuracy:** - 0,996

	Precision	Recall	F1-Score	Support
da	0.996	1	0.998	844
de	0.999	0.997	0.998	1448
en	0.997	0.997	0.997	2519
fr	0.927	1	0.962	38
it	0.833	0.781	0.806	32
accuracy			0.996	4881
macro avg	0.951	0.955	0.952	4881
weighted avg	0.996	0.996	0.996	4881



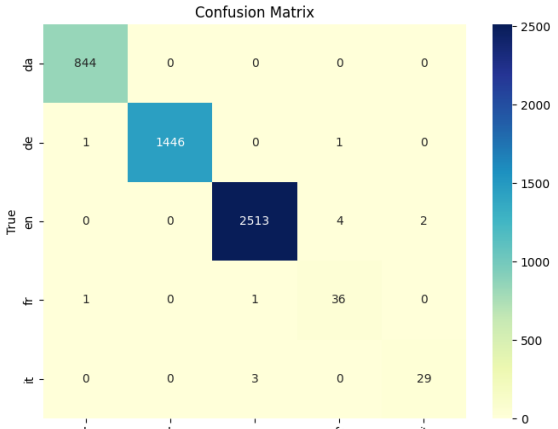
Imlementierung & Evaluierung: Sprachklassifizierung

Klassifikation nach Sprache mit **k=5**

Classification Report:

- **Accuracy:** - 0,997

	Precision	Recall	F1-Score	Support
da	0.998	1	0.999	844
de	1	0.999	0.999	1448
en	0.998	0.998	0.998	2519
fr	0.878	0.947	0.911	38
it	0.935	0.906	0.921	32
accuracy			0.997	4881
macro avg	0.962	0.97	0.966	4881
weighted avg	0.997	0.997	0.997	4881



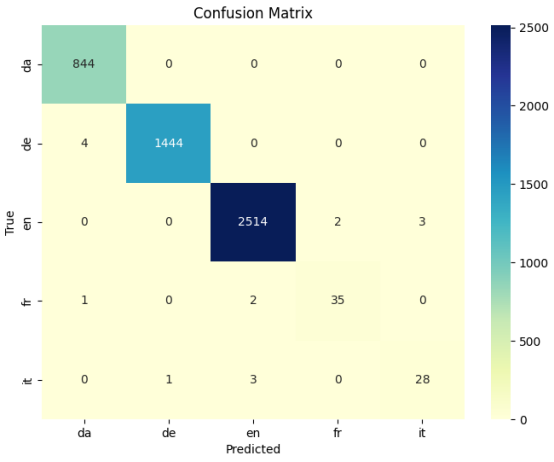
Imlementierung & Evaluierung: Sprachklassifizierung

Klassifikation nach Sprache mit **k=257**

Classification Report:

- **Accuracy:** - 0,997

	Precision	Recall	F1-Score	Support
da	0.994	1	0.997	844
de	0.999	0.997	0.998	1448
en	0.998	0.998	0.998	2519
fr	0.946	0.921	0.933	38
it	0.903	0.875	0.889	32
accuracy			0.997	4881
macro avg	0.968	0.958	0.963	4881
weighted avg	0.997	0.997	0.997	4881



Ergebnis Sprachklassifizierung:

1. k=1:
   - Sehr hohe Genauigkeit (accuracy) von 99,6
   - Hohe Präzision, Recall und F1-Score für die meisten Sprachen.
   - Geringe Werte für Französisch (fr) und Italienisch (it).
2. k=5:
   - Minimale Verbesserung der Genauigkeit auf 99,7
   - Verbesserung in den Werten für Französisch (fr) und Italienisch (it) im Vergleich zu k=1.
3. k=257:
   - Weiterhin hohe Genauigkeit von 99,7
   - Gute Präzision und Recall für alle Sprachen im Vergleich zu k=5 und k=1.

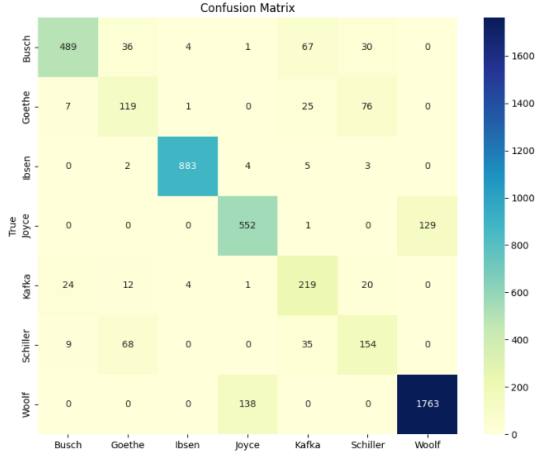
Imlementierung & Evaluierung: Autorenklassifizierung

Klassifikation nach Autoren und Autorinnen mit **k=1**

Classification Report:

- **Accuracy:** - 0,856

	Precision	Recall	F1-Score	Support
Busch	0.924	0.78	0.846	627
Goethe	0.502	0.522	0.512	228
Ibsen	0.99	0.984	0.987	897
Joyce	0.793	0.809	0.801	682
Kafka	0.622	0.782	0.693	280
Schiller	0.544	0.579	0.561	266
Woolf	0.932	0.927	0.93	1901
accuracy			0.856	4881
macro avg	0.758	0.769	0.761	4881
weighted avg	0.863	0.856	0.858	4881



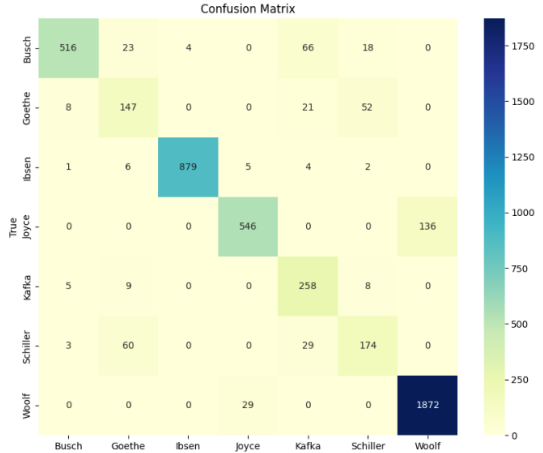
Imlementierung & Evaluierung: Autorenklassifizierung

Klassifikation nach Autoren und Autorinnen mit **k=13**

Classification Report:

- **Accuracy:** - 0,900

	Precision	Recall	F1-Score	Support
Busch	0.968	0.823	0.89	627
Goethe	0.6	0.645	0.622	228
Ibsen	0.995	0.98	0.988	897
Joyce	0.941	0.801	0.865	682
Kafka	0.683	0.921	0.784	280
Schiller	0.685	0.654	0.669	266
Woolf	0.932	0.985	0.958	1901
accuracy			0.9	4881
macro avg	0.829	0.83	0.825	4881
weighted avg	0.906	0.9	0.9	4881

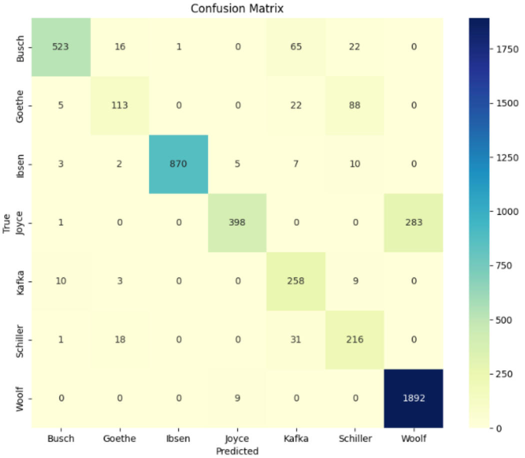


Klassifikation nach Autoren und Autorinnen mit **k=257**

Classification Report:

- **Accuracy:** - 0,875

	Precision	Recall	F1-Score	Support
Busch	0.963	0.834	0.894	627
Goethe	0.743	0.496	0.595	228
Ibsen	0.999	0.97	0.984	897
Joyce	0.966	0.584	0.728	682
Kafka	0.674	0.921	0.778	280
Schiller	0.626	0.812	0.707	266
Woolf	0.87	0.995	0.928	1901
accuracy			0.875	4881
macro avg	0.834	0.802	0.802	4881
weighted avg	0.889	0.875	0.87	4881



Ergebnis Autorenklassifizierung:

1. k=1:
   - Genauigkeit von 85,6
   - Gute Leistung für Autoren wie Busch, Ibsen und Woolf.
   - Schlechtere Leistung für Autoren wie Goethe und Schiller.
2. k=13:
   - Deutliche Verbesserung der Genauigkeit auf 89,9
   - Gute Precision für die meisten Autoren, insbesondere Verbesserungen für Joyce und leice Verbesserung für Schiller.
3. k=257:
   - Leichtes sinken der Accuracy auf 87,5
   - Ähnliche Precision wie bei k=13, nur leichtes sinken für Woolf.

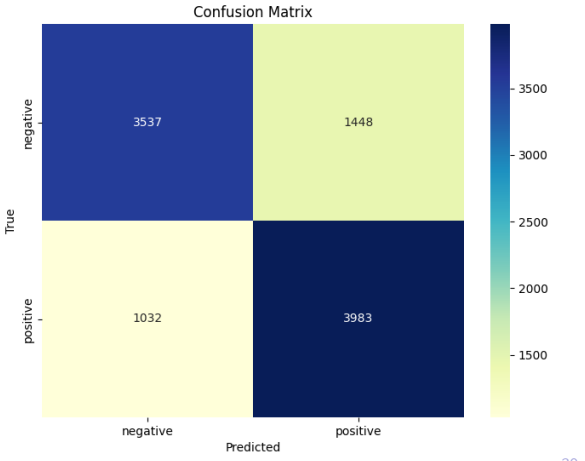
Imlementierung & Evaluierung: Sentimentklassifizierung

Klassifikation nach Sentiment mit **k=1**

Classification Report:

- **Accuracy:** - 0,75

	Precision	Recall	F1-Score	Support
negative	0.77	0.71	0.74	4985
positive	0.73	0.79	0.76	5015
accuracy			0.75	10000
macro avg	0.75	0.75	0.75	10000
weighted avg	0.75	0.75	0.75	10000



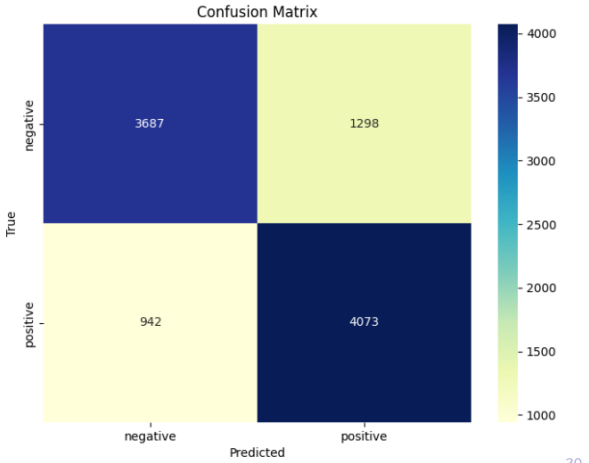
Imlementierung & Evaluierung: Sentimentklassifizierung

Klassifikation nach Sentiment mit **k=7**

Classification Report:

- **Accuracy:** - 0,78

	Precision	Recall	F1-Score	Support
negative	0.8	0.74	0.77	4985
positive	0.76	0.81	0.78	5015
accuracy			0.78	10000
macro avg	0.78	0.78	0.78	10000
weighted avg	0.78	0.78	0.78	10000



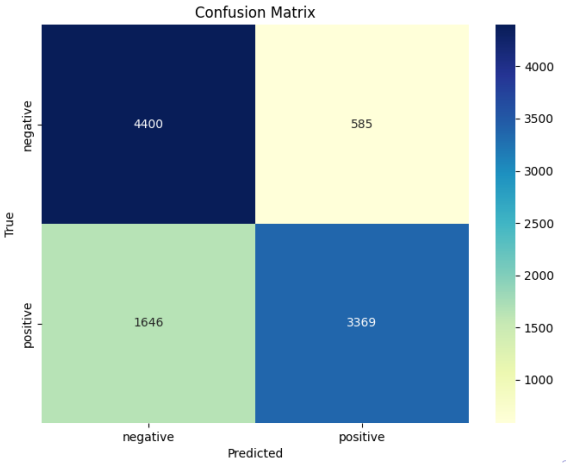
Imlementierung & Evaluierung: Sentimentklassifizierung

Klassifikation nach Sentiment mit **k=257**

Classification Report:

- **Accuracy:** - 0,78

	Precision	Recall	F1-Score	Support
negative	0.73	0.88	0.8	4985
positive	0.85	0.67	0.75	5015
accuracy			0.78	10000
macro avg	0.79	0.78	0.77	10000
weighted avg	0.79	0.78	0.77	10000



Ergebnis Sentimentklassifizierung:

1. k=1:
   - Genauigkeit von 75,2
2. k=7:
   - leichtes Anstieg der Accuracy.
3. k=257:
   - Ähnliche Genauigkeit wie bei k=257.
   - Bessere Leistung im Recall für negatives Sentiment und bei Precision für positives Sentiment.

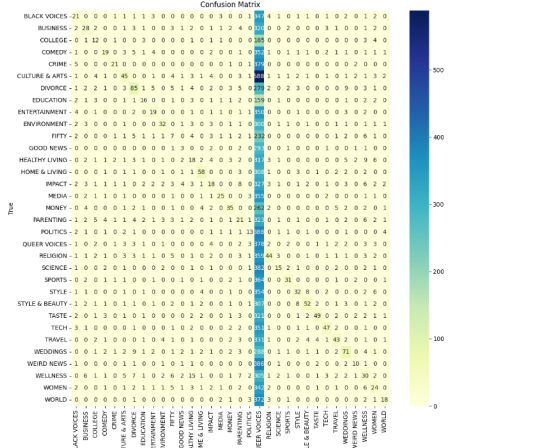
Imlementierung & Evaluierung: Newsklassifizierung

Klassifikation nach News mit **k=1**

Classification Report:

- **Accuracy:** - 0,104

	Precision	Recall	F1-Score	Support
BLACK VOICES	0.356	0.054	0.093	392
BUSINESS	0.452	0.074	0.127	380
COLLEGE	0.286	0.057	0.094	212
COMEDY	0.442	0.048	0.086	399
CRIME	0.677	0.051	0.095	409
CULTURE & ARTS	0.536	0.097	0.119	673
DIVORCE	0.603	0.203	0.304	419
EDUCATION	0.421	0.081	0.136	198
ENTERTAINMENT	0.452	0.049	0.089	367
ENVIRONMENT	0.827	0.09	0.158	355
FIFTY	0.14	0.026	0.043	273
GOOD NEWS	0.136	0.01	0.018	305
HEALTHY LIVING	0.247	0.047	0.078	386
HOME & LIVING	0.725	0.15	0.249	386
IMPACT	0.316	0.045	0.079	400
MEDIA	0.676	0.063	0.116	395
MONEY	0.595	0.106	0.181	324
PARENTING	0.256	0.059	0.089	391
POLITICS	0.394	0.031	0.057	420
QUEER VOICES	0.035	0.911	0.067	415
RELIGION	0.62	0.1	0.172	440
SCIENCE	0.5	0.036	0.068	414
SPORTS	0.66	0.076	0.136	410
STYLE	0.552	0.077	0.136	413
STYLE & BEAUTY	0.684	0.133	0.223	391
TASTE	0.79	0.123	0.214	367
TECH	0.701	0.113	0.195	418
TRAVEL	0.614	0.106	0.181	405
WEDDINGS	0.802	0.179	0.274	409
WEIRD NEWS	0.455	0.025	0.047	408
WELLNESS	0.283	0.074	0.117	407
WOMEN	0.338	0.06	0.102	400
WORLD	0.545	0.045	0.082	404
accuracy			0.104	11924
macro avg	0.475	0.102	0.128	11924
weighted avg	0.486	0.104	0.129	11924



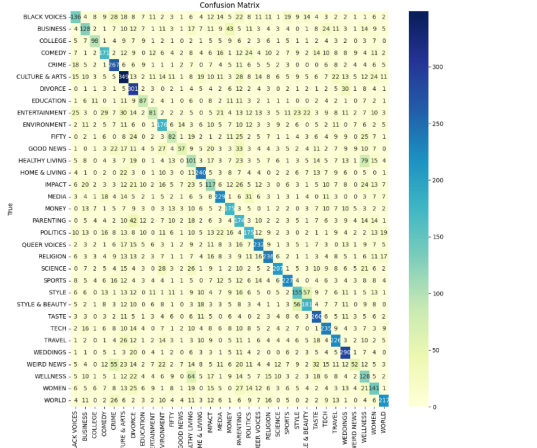
Imlementierung & Evaluierung: Newsklassifizierung

Klassifikation nach News mit **k=77**

Classification Report:

- **Accuracy:** - 0,462

	Precision	Recall	F1-Score	Support
BLACK VOICES	0.459	0.347	0.395	392
BUSINESS	0.4	0.337	0.366	380
COLLEGE	0.594	0.462	0.52	212
COMEDY	0.475	0.429	0.451	399
CRIME	0.513	0.653	0.575	409
CULTURE & ARTS	0.481	0.519	0.498	673
DIVORCE	0.431	0.718	0.538	419
EDUCATION	0.435	0.439	0.437	198
ENTERTAINMENT	0.407	0.209	0.276	367
ENVIRONMENT	0.419	0.496	0.454	355
FIFTY	0.364	0.3	0.329	273
GOOD NEWS	0.483	0.187	0.27	305
HEALTHY LIVING	0.217	0.262	0.237	386
HOME & LIVING	0.603	0.622	0.612	386
IMPACT	0.297	0.293	0.295	400
MEDIA	0.531	0.58	0.554	395
MONEY	0.44	0.54	0.485	324
PARENTING	0.288	0.445	0.349	391
POLITICS	0.431	0.417	0.424	420
QUEER VOICES	0.526	0.559	0.542	415
RELIGION	0.602	0.536	0.567	440
SCIENCE	0.663	0.5	0.57	414
SPORTS	0.624	0.554	0.587	410
STYLE	0.449	0.375	0.409	413
STYLE & BEAUTY	0.529	0.483	0.494	391
TASTE	0.537	0.655	0.59	397
TECH	0.527	0.565	0.545	418
TRAVEL	0.499	0.558	0.527	405
WEDDINGS	0.549	0.725	0.625	409
WEIRD NEWS	0.403	0.127	0.194	408
WELLNESS	0.285	0.314	0.298	407
WOMEN	0.367	0.352	0.36	400
WORLD	0.829	0.537	0.579	404
accuracy			0.462	12824
macro avg	0.467	0.457	0.453	12824
weighted avg	0.468	0.462	0.456	12824



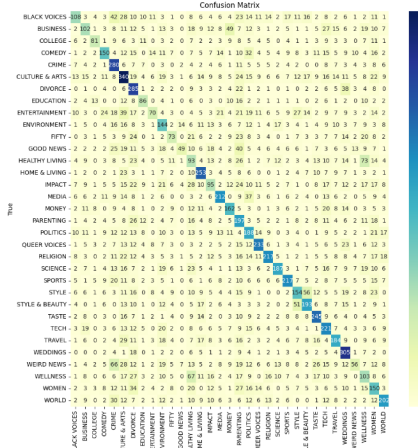
Imlementierung & Evaluierung: Newsklassifizierung

Klassifikation nach News mit **k=257**

Classification Report:

- **Accuracy:** - 0,439

	Precision	Recall	F1-Score	Support
BLACK VOICES	0.489	0.276	0.352	392
BUSINESS	0.388	0.288	0.331	380
COLLEGE	0.587	0.382	0.463	212
COMEDY	0.497	0.376	0.428	399
CRIME	0.435	0.685	0.532	409
CULTURE & ARTS	0.426	0.505	0.462	673
DIVORCE	0.399	0.68	0.503	419
EDUCATION	0.434	0.434	0.434	198
ENTERTAINMENT	0.459	0.181	0.259	367
ENVIRONMENT	0.345	0.406	0.373	355
FIFTY	0.271	0.207	0.311	273
GOOD NEWS	0.805	0.161	0.254	305
HEALTHY LIVING	0.197	0.241	0.217	386
HOME & LIVING	0.549	0.655	0.597	386
IMPACT	0.31	0.238	0.269	400
MEDIA	0.568	0.537	0.552	395
MONEY	0.453	0.5	0.475	324
PARENTING	0.294	0.504	0.371	391
POLITICS	0.373	0.448	0.407	420
QUEER VOICES	0.594	0.561	0.577	415
RELIGION	0.583	0.493	0.534	440
SCIENCE	0.623	0.452	0.524	414
SPORTS	0.59	0.529	0.558	410
STYLE	0.422	0.373	0.396	413
STYLE & BEAUTY	0.519	0.494	0.506	391
TASTE	0.556	0.617	0.585	367
TECH	0.499	0.531	0.462	418
TRAVEL	0.468	0.454	0.461	405
WEDDINGS	0.938	0.762	0.81	409
WEIRD NEWS	0.397	0.137	0.204	408
WELLNESS	0.253	0.253	0.253	407
WOMEN	0.304	0.375	0.336	400
WORLD	0.571	0.5	0.533	404
accuracy			0.439	12824
macro avg	0.453	0.433	0.428	12824
weighted avg	0.452	0.439	0.431	12824



Ergebnis Newsklassifizierung:

1. k=1:
   - Geringe Genauigkeit von 10,4%.
   - Herausforderungen bei der Unterscheidung zwischen verschiedenen Genres.
2. k=77:
   - Verbesserte Genauigkeit auf 46,2%.
   - Bessere Leistung für bestimmte Genres wie DIVORCE, POLITICS und WEDDINGS.
3. k=257:
   - Weiterhing nicht sehr verschlechtete Genauigkeit auf 43,9%.

Zusammenfassung

1. Die Sprachklassifizierung erreicht eine hohe Genauigkeit mit dem KNN-Algorithmus, wobei k=5 eine gute Balance zwischen Precision und Recall zu bieten scheint.
2. Die Autorenklassifizierung profitiert von höheren k-Werten, wobei k=13 die beste Gesamtleistung zeigt.
   - **Textinhalt und Merkmale:** Die Art des Textinhalts bei der Sprach- und Autorenklassifikationen kann auf sprachlichen Mustern und Schreibstilen basieren, die für das Modell leichter zu lernen sind.
3. Bei der Sentimentklassifikation sind die Accuracy Werte im Vergleich ein Bisschen niedriger = 0,78
   - Hier kann **der Umfang des Trainingsdatensatzes** eine wichtige Rolle spielen. Auch **die Art des Textes** kann ebenfalls einen Einfluss haben. Sentimentanalyse erfordert eine differenzierte Interpretation von Texten.

Zusammenfassung

4. Die Newsklassifizierung bleibt eine Herausforderung, und höhere k-Werte führen zu einer leichten Verbesserung, obwohl die Genauigkeit insgesamt niedrig bleibt.
   - Es gibt hier **eine Vielzahl von Klassen**, was die Aufgabe komplexer macht. Zudem könnte die Verteilung der Beispiele auf die Klassen ungleichmäßig sein, was die Modellgenauigkeit beeinflussen kann.
   - **Textinhalt und Merkmale:** Die Art des Textinhalts kann stark variieren. In den News-Klassifizierungsaufgaben müssen Modelle möglicherweise subtile semantische Unterschiede zwischen verschiedenen Artikeln erfassen.

Literaturquellen

- Antonio Mucherino, Petraq J. Papajorgji Panos M. Pardalos. *k-Nearest Neighbor Classification.* Data Mining in Agriculture. pp 83–106 (2009)
- Auliya Rahman Isnain, Jepi Supriyanto, Muhammad Pajar Kharisma *Implementation of K-Nearest Neighbor (K-NN) AlgorithmFor Public Sentiment Analysis of Online Learning.* Indonesian Journal of Computing and Cybernetics Systems. IJCCS Vol. 15, No.2, April2021: 121–130
- E.A. Patrick, F.P. Fischer III. *A generalized k-nearest neighbor rule.*. Information and Control. Volume 16, Issue 2, April 1970, Pages 128-152
- Zhongheng Zhang. *Introduction to machine learning: k-nearest neighbors* Ann Transl Med. 2016 Jun; 4(11): 218

Haben Sie noch Fragen?

Vielen Dank für Eure Aufmerksamkeit!