

Support Vector Machine (SVM)

Ludwig-Maximilians-Universität, München

Centrum für Informations- und Sprachverarbeitung

Masterseminar: Klassifikation und Clustering

Dozent: PD Dr. Stefan Langer

Referentin: Evanthia Tsigkana

München, 11. Dezember 2023



Übersicht

1. Einführung
2. Konzept des Algorithmus SVM
3. Vor- und Nachteile
4. Implementierung & Evaluation
5. Literatur

Einführung

Zur geschichtlichen Entwicklung des Algorithmus:

- 1936: R.A. Fisher veröffentlichte die Idee der Trennung durch eine Hyperebene
- 1958: Rosenblatt veröffentlichte eine ähnliche Idee zum Thema Support Vector Machines
- 1992: Wapnik, Boser, Guyon verwendeten den Kernel-Trick
- Anfang der 1990er Jahre beschäftigte sich Corinna Cortes mit SVM
- Mitte der 1990er Jahre Durchbruch der SVMs

Einführung

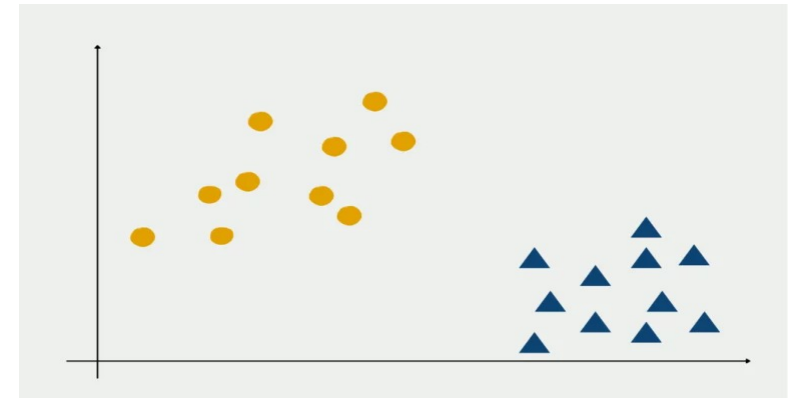
Was sind die Support Vector Machines (SVMs)?

- Mathematische Algorithmen in Machine Learning genutzt → Klassifikation von Objekten – supervised learning
- Bei Text- und Bildklassifizierung vorteilhafter gegenüber Neuronale Netzwerke → schnell trainierbar, wenige Trainingsdaten erforderlich, gute Ergebnisse
- Es gibt lineare SVMs, nicht-lineare SVMs und Support Vector Regression. (Die SVR versucht den Abstand zwischen den vorhergesagten Ausgaben und den tatsächlichen Ausgaben zu minimieren.)

Konzept

Wie funktioniert der Algorithmus?

- Die SVM entscheidet sich, ob das Objekt als blau oder gelb klassifiziert wird.
- Trainingsdaten → in Koordinatensystem abgebildet, weil wir nur 2 Dimensionen haben

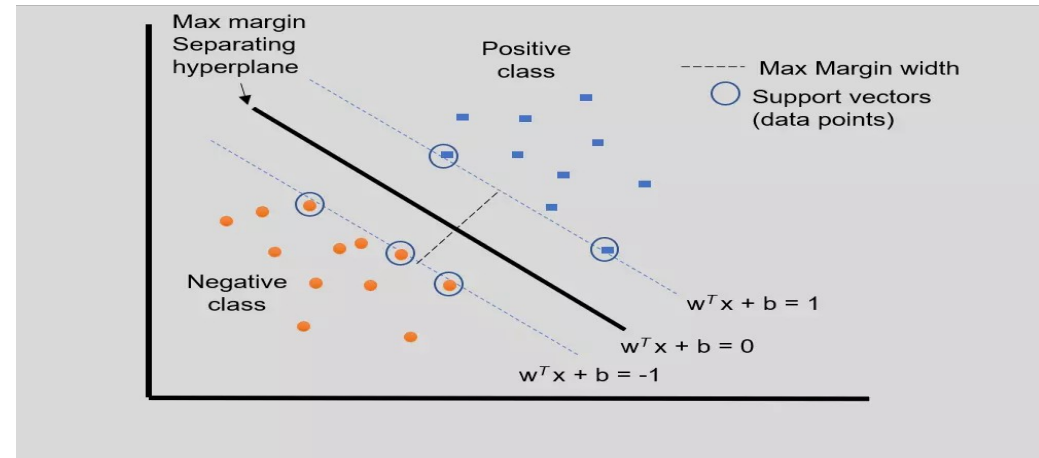


(Abb1., Quelle: <https://databasecamp.de/ki/support-vector-machine-svm>. Zugriff am 08.12.2023)

Konzept

- Die SVM trennt die Klassen durch Hyperebenen im Vektorraum; im zweidimensionalen Raum → eine einfache Linie
- Der Normalvektor steht senkrecht auf allen Vektoren der Hyperebene
- Es gilt:

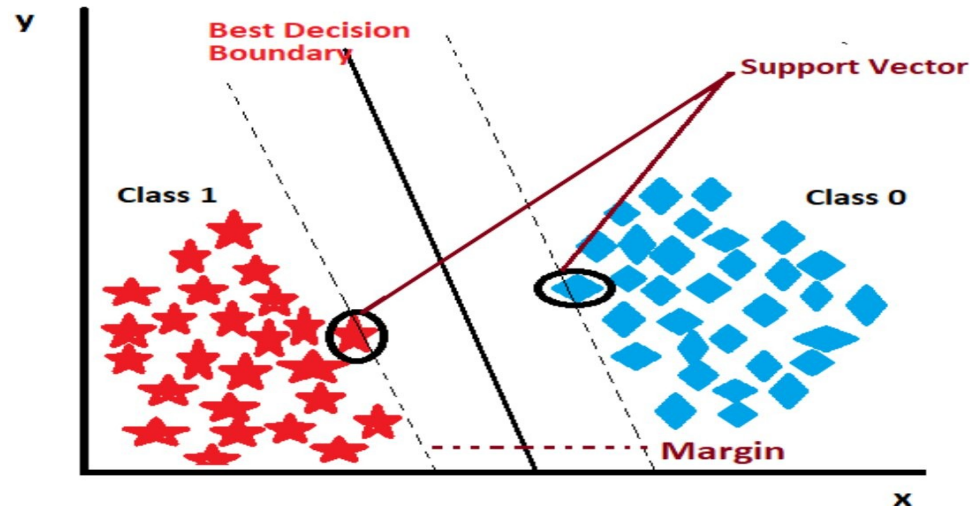
$$w^*x + b \begin{cases} > 0 \text{ falls } x \text{ im positiven Raum} \\ = 0 \text{ falls } x \text{ auf } H \\ < 0 \text{ falls } x \text{ im negativen Raum} \end{cases}$$



(Abb. 2, 3, Quelle: <https://nguyenkm.com/projects/machine-learning/support-vector-machines/>. Zugriff am 08.12.2023)

Konzept

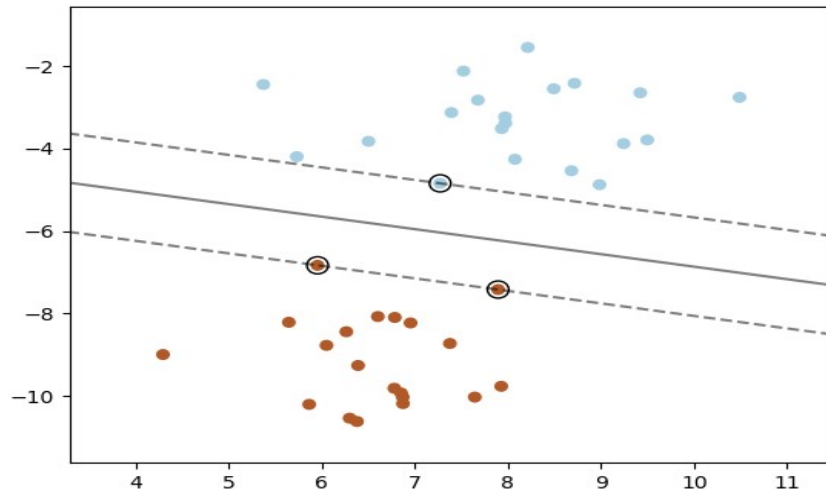
- Mehrere Trainings $\rightarrow$ bis der Abstand (Margin) zwischen den Elementen jeder Gruppe zu der Hyperebene maximal wird.
- Die SVM trennt beide Klassen möglichst stark.



(Abb.4, Quelle: <https://statusneo.com/defying-convention-svm-the-maverick-of-ml-algorithms>. Zugriff am 08.12.2023)

Konzept

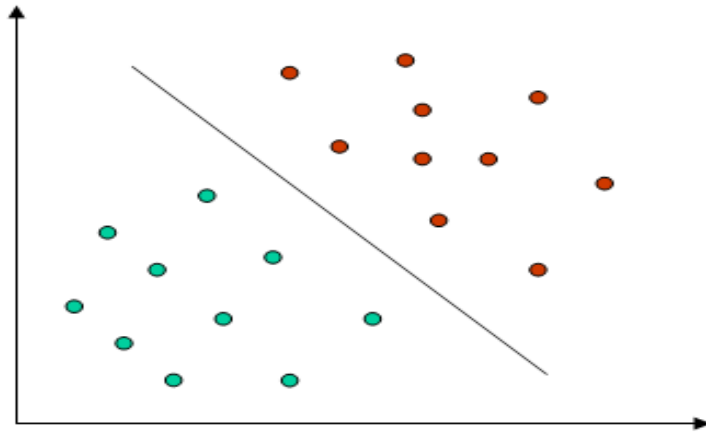
Vektoren nahe der Hyperebene haben größeren Einfluss →
Stützvektoren (support vectors)



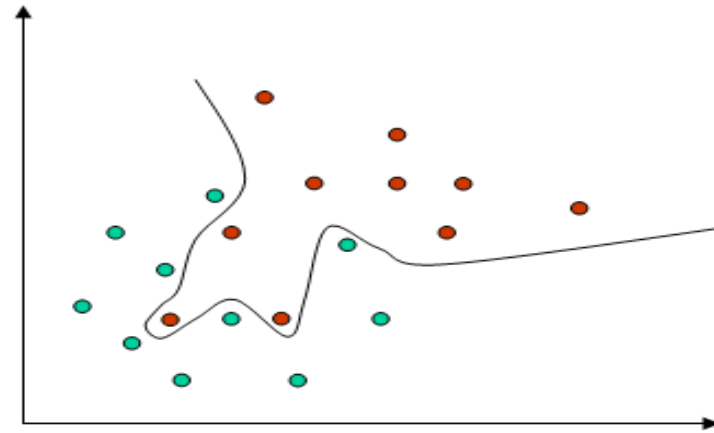
(Abb. 5, Quelle: https://scikit-learn.org/stable/_images/sphx_glr_plot_separating_hyperplane_001.png. Zugriff am 08.12.2023)

Konzept

Aber: Nicht alle Probleme linear trennbar



linear trennbar



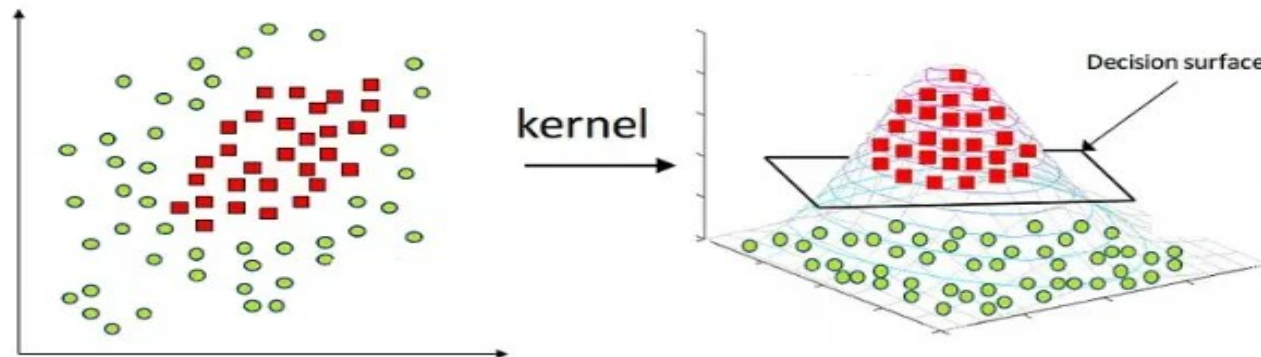
nicht linear trennbar

(Abb. 6, 7, Quelle: https://de.wikipedia.org/wiki/Support_Vector_Machine#/media/Datei:Diskriminanzfunktion.png. Zugriff am 08.12.2023)

Konzept

Mögliche Lösung:

- Andere Hyperflächen finden durch den Kernel - Trick: Transformieren der Datenmengen in einen höherdimensionalen Raum, in dem lineare Trennung möglich ist, dann Rücktransformation



(Abb. 8, 9, Quelle: https://medium.com/@Suraj_Yadav/what-is-kernel-trick-in-svm-interview-questions-related-to-kernel-trick-97674401c48d.)

Zugriff am 08.12.2023)

Kernel-Funktionen:

- Lineare Kernel: $k(x, y) = \langle x, y \rangle$
 - Er wird benutzt, wenn die Daten einfach trennbar sind.
 - Sehr einfach und funktioniert gut bei nicht höher-dimensionalen Datensätzen mit einer großen Anzahl an Merkmalen
- Polynomiale Kernel: $k(x, y) = \langle x, y \rangle^d$, mit einem freien Parameter d
 - Er bildet die Eingabedaten in einem höher-dimensionalen Raum
- Radiale-Basisfunktion-Kernel (RBF): $k(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right)$
 - Er wird benutzt, wenn die Daten nicht trennbar durch eine lineare oder polynome Hyperbene sind.

Vor- und Nachteile von SVM

Vorteile

- Modelltraining ist relativ einfach
- Geeignet für höherdimensionale Daten
- Nützlich sowohl für lineare als auch für nicht-linear trennbare Daten
- Kann alternativ zum Neuralen Netzwerk verwendet werden

Nachteile

- Große Mengen an Datensätzen werden benötigt
- Noise kann nicht sehr gut gefiltert werden
- Die Ergebnisse werden durch die Hyperebene entschieden – es besteht keine Wahrscheinlichkeitsinterpretation
- Hohe Computerleistung für die Bestimmung der Kernelfunktion

Implementierung & Evaluation

Briefsammlung – Kategorisierung nach Sprache

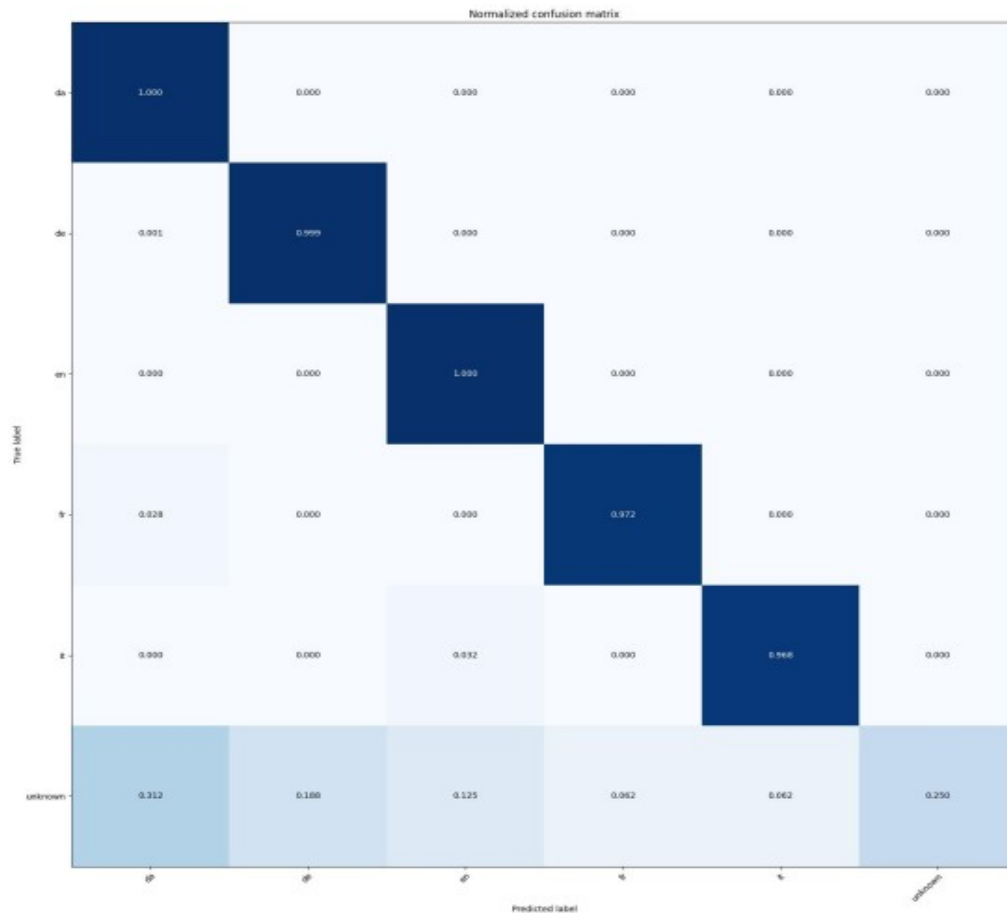
Methode: wortbasierte TF-IDF-Matrix

Classifier: LinearSVC-dense

	Precision	Recall	f1-score	Support
da	0.992	1.000	0.996	838
de	0.998	0.999	0.999	1443
en	0.999	1.000	0.999	2517
fr	0.972	0.972	0.972	36
it	0.938	0.968	0.952	31
unk.	1.000	0.250	0.400	16
accuracy			0.997	4881
macro avg.	0.983	0.865	0.886	4881
weight avg	0.997	0.997	0.996	4881

f1 score: de 0,999, en 0,999, da 0,996

accuracy: 0,997



Implementierung & Evaluation

Briefsammlung – Kategorisierung nach Sprache

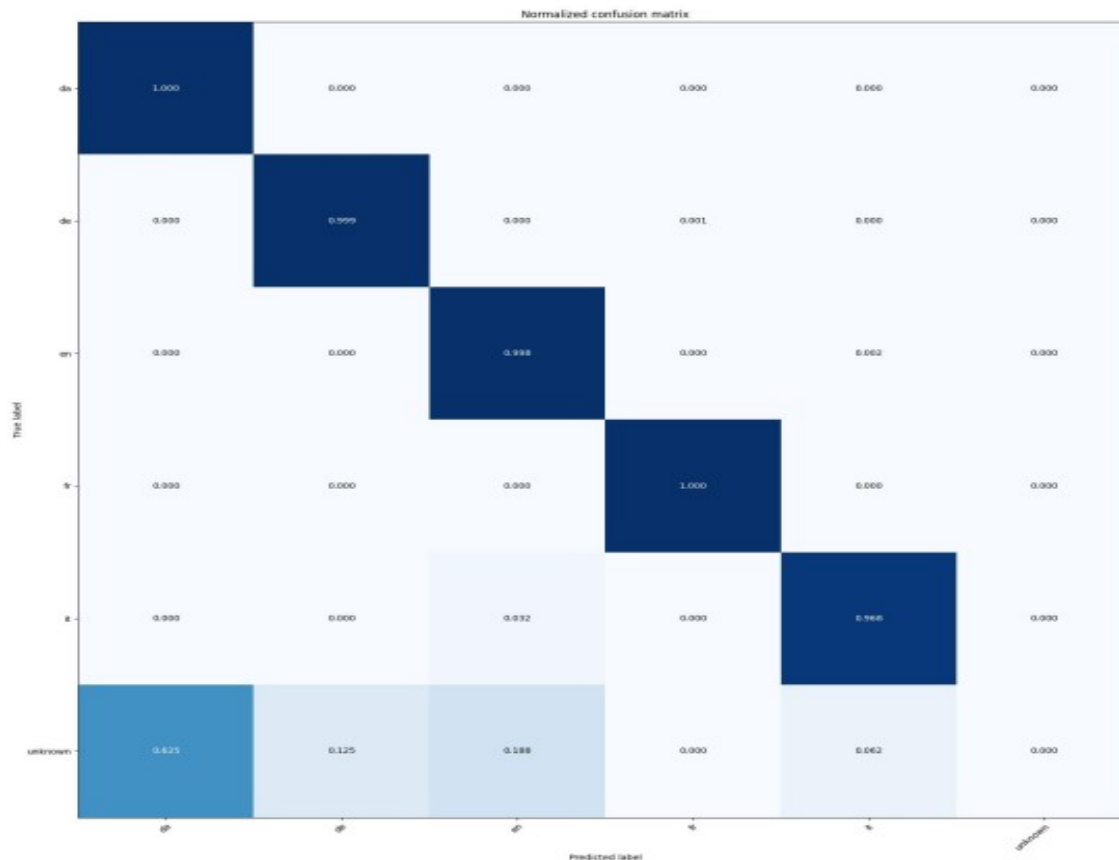
Methode: wortbasierte TF-IDF-Matrix

Classifier: LinearSVC-LSA

	Precision	Recall	f1-score	Support
da	0.998	1.000	0.994	838
de	0.999	0.999	0.999	1443
en	0.998	0.998	0.998	2517
fr	0.947	1.000	0.973	36
it	0.857	0.968	0.909	31
Unk.	0.000	0.000	0.000	16
accuracy			0.995	4881
macro avg.	0.798	0.828	0.812	4881
weight avg	0.992	0.995	0.994	4881

f1-score: de 0,999, en 0,998, da 0,994

Accuracy: 0,995



Implementierung & Evaluation

Briefsammlung – Kategorisierung nach Sprache

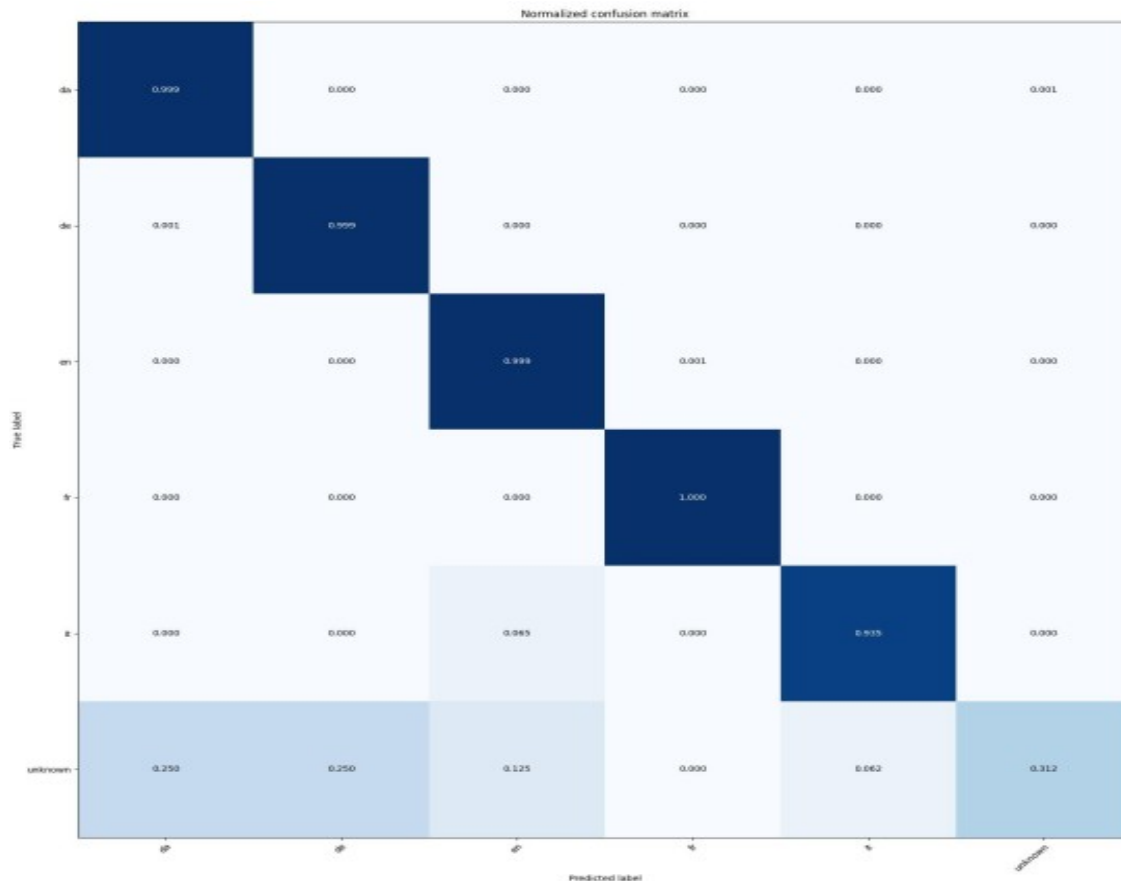
Methode: wortbasierte TF-IDF-Matrix

Classifier: LinearSVC-sparse

	Precision	Recall	f1-score	Support
da	0.994	0.999	0.996	838
de	0.997	0.999	0.998	1443
en	0.998	0.999	0.999	2517
fr	0.947	1.000	0.973	36
it	0.935	0.935	0.935	31
unk.	0.833	0.312	0.455	16
accuracy			0.996	4881
macro avg.	0.951	0.874	0.893	4881
weight avg	0.996	0.996	0.996	4881

f1-score: en 0,999, de 0,998, da 0,996

Accuracy: 0,996



Implementierung & Evaluation

Briefsammlung – Kategorisierung nach Sprache

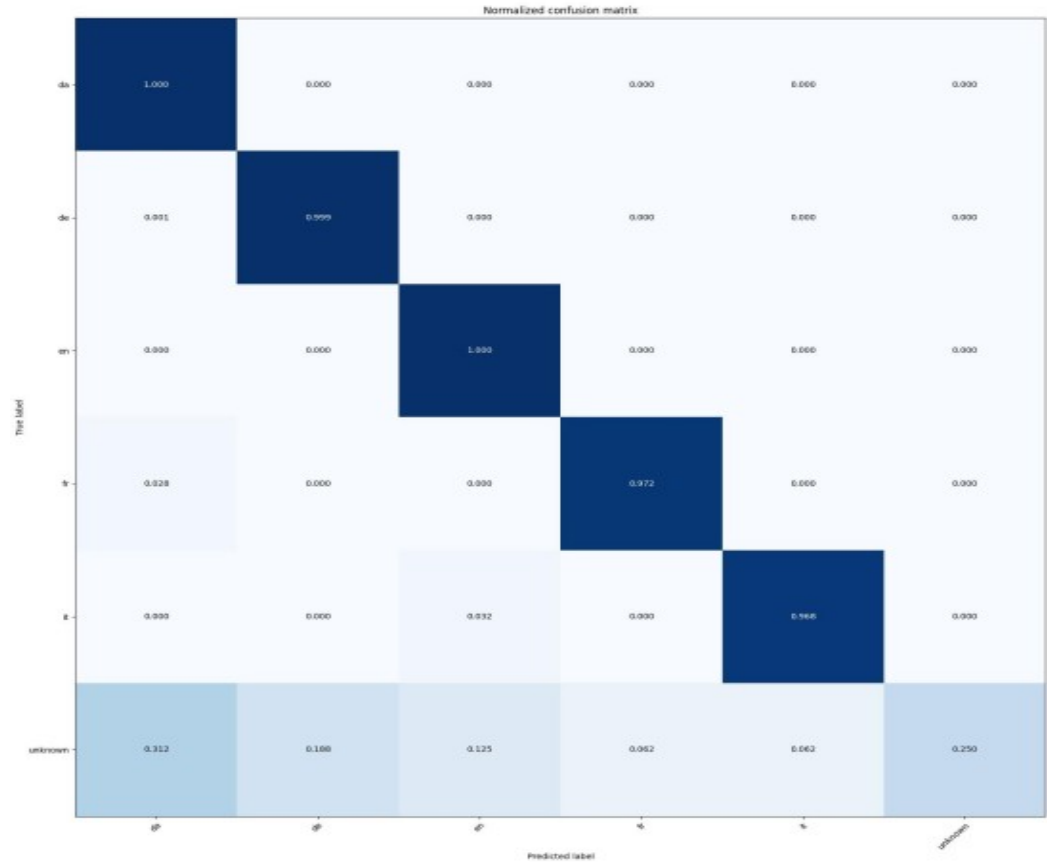
Methode: Bigramme TF-IDF-Matrix

Classifier: LinearSVC-dense

	Precision	Recall	f1-score	Support
da	0.992	1.000	0.996	838
de	0.998	0.999	0.999	1443
en	0.999	1.000	0.999	2517
fr	0.972	0.972	0.972	36
it	0.938	0.968	0.952	31
Unk.	1.000	0.250	0.400	16
accuracy			0.997	4881
macro avg.	0.983	0.865	0.886	4881
weight avg	0.997	0.997	0.996	4881

f1-score: en 0,999, de 0,999, da 0,996

Accuracy: 0,997



Implementierung & Evaluation

Briefsammlung – Kategorisierung nach Sprache

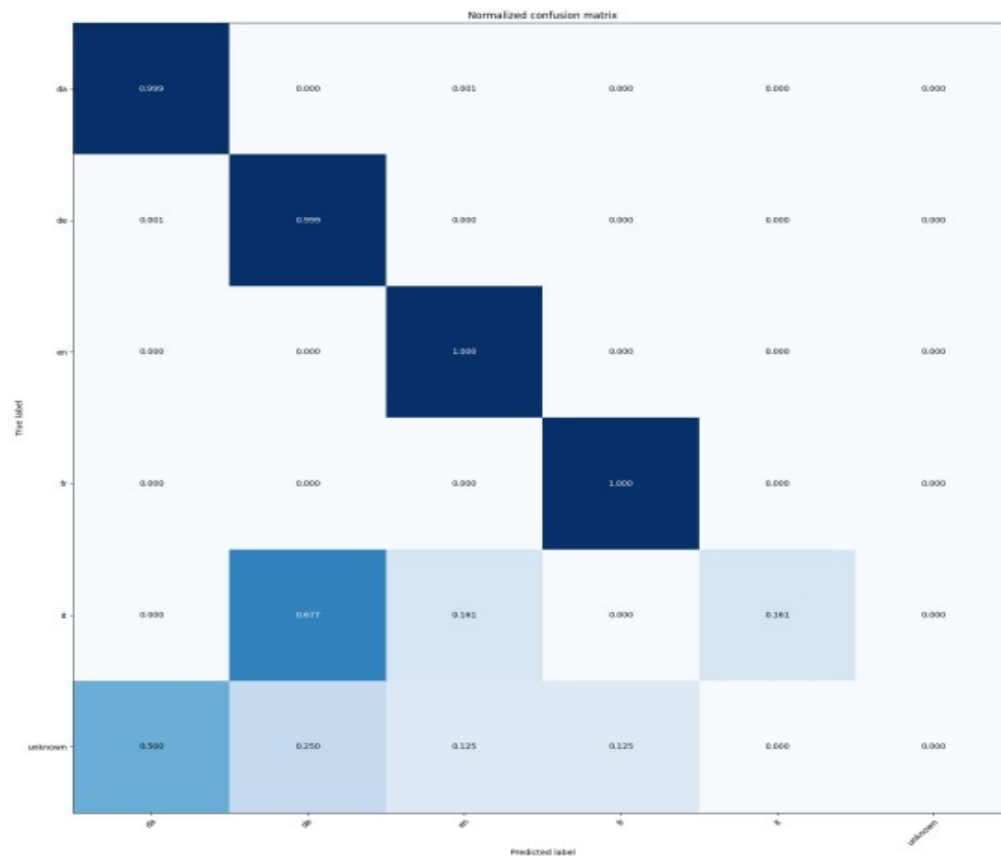
Methode: Bigramme TF-IDF-Matrix

Classifier: LinearSVC-LSA

	Precision	Recall	f1-score	Support
da	0.989	0.999	0.994	838
de	0.985	0.999	0.992	1443
en	0.994	1.000	0.997	2517
fr	0.947	1.000	0.973	36
it	1.000	0.032	0.062	31
Unk.	0.000	0.000	0.000	16
accuracy			0.999	4881
macro avg.	0.819	0.672	0.670	4881
weight avg	0.987	0.990	0.985	4881

f1-score: en 0,997, de 0,992, da 0,994

Accuracy: 0,990



Implementierung & Evaluation

Briefsammlung – Kategorisierung nach Autorinnen und Autoren

Methode: wortbasierte TF-IDF-Matrix

Classifier: LinearSVC-dense

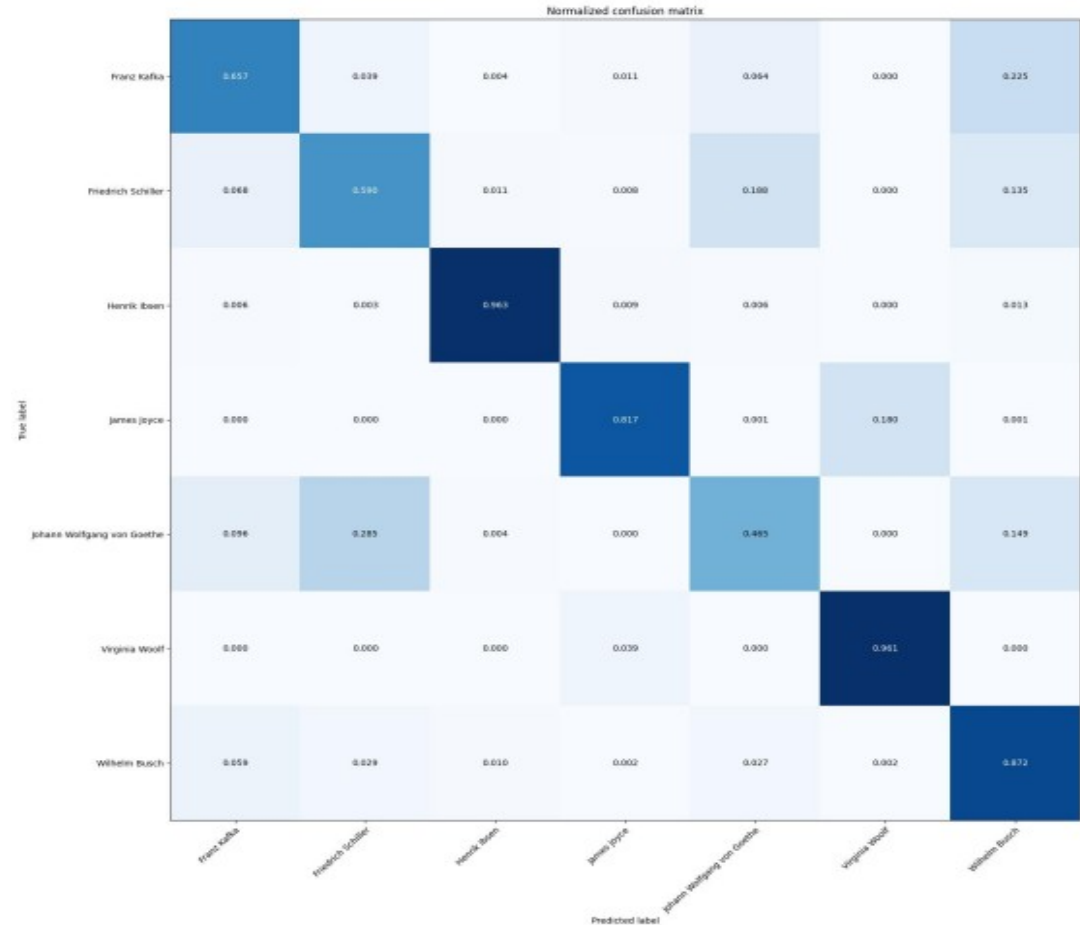
	precision	recall	f1-score	support
Kafka	0.692	0.657	0.674	280
Schiller	0.618	0.590	0.604	266
Ibsen	0.987	0.963	0.975	897
Joyce	0.862	0.817	0.839	682
Goethe	0.538	0.465	0.499	228
Woolf	0.936	0.961	0.948	1901
Busch	0.789	0.872	0.829	627

accuracy			0.869	4881
macro avg	0.775	0.761	0.767	4881
weighted avg	0.867	0.869	0.867	4881

Ergebnisse:

f1-score: Ibsen 0,975; Woolf 0,948; Joyce 0,839

accuracy: 0,869



Implementierung & Evaluation

Briefsammlung – Kategorisierung nach Autorinnen und Autoren

Methode: wortbasierte TF-IDF-Matrix

Classifier: LinearSVC-LSA

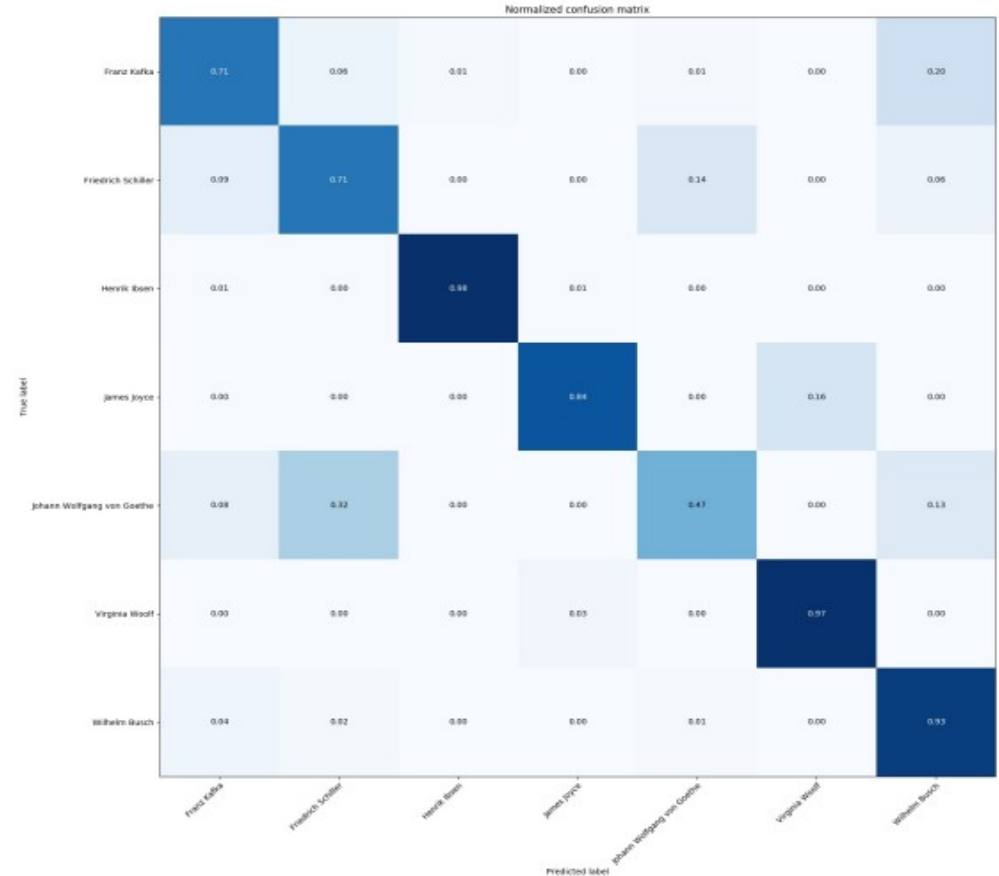
	Precision	recall	f1-score	support
Kafka	0.745	0.711	0.728	280
Schiller	0.630	0.703	0.664	266
Ibsen	0.994	0.975	0.985	897
Joyce	0.905	0.849	0.876	682
Goethe	0.675	0.491	0.569	228
Woolf	0.948	0.971	0.959	1901
Busch	0.854	0.930	0.890	627

accuracy			0.898	4881
macro avg	0.821	0.804	0.810	4881
weighted avg	0.896	0.898	0.896	4881

Ergebnisse:

f1-score: Ibsen 0,985, Woolf 0,959, Busch 0,890

accuracy: 0,898



Implementierung & Evaluation

Briefsammlung – Kategorisierung nach Autorinnen und Autoren

Methode: wortbasierte TF-IDF-Matrix

Classifier: LinearSVC-sparse

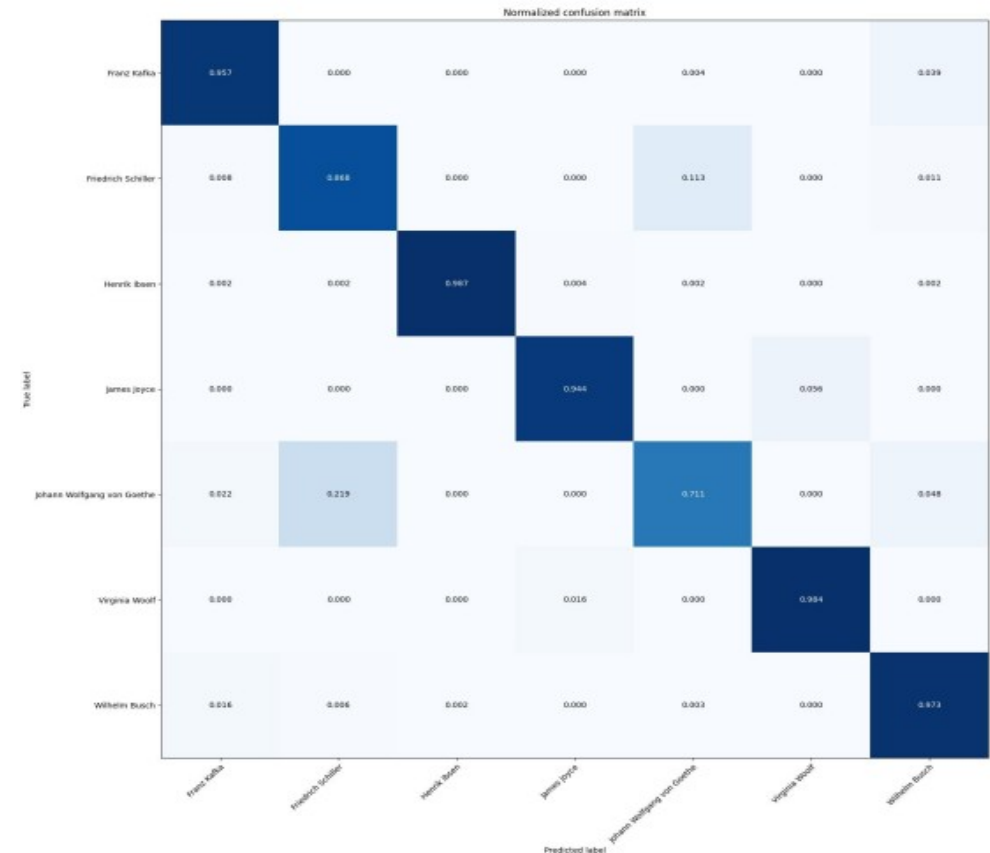
	Precision	recall	f1-score	support
Kafka	0.934	0.957	0.945	280
Schiller	0.805	0.868	0.835	266
Ibsen	0.999	0.987	0.993	897
Joyce	0.948	0.944	0.946	682
Goethe	0.822	0.711	0.762	228
Woolf	0.980	0.984	0.982	1901
Busch	0.958	0.973	0.965	627

accuracy			0.957	4881
macro avg	0.921	0.918	0.918	4881
weighted avg	0.957	0.957	0.956	4881

Ergebnisse:

f1-score: Ibsen 0,993, Woolf 0,982, Busch 0,965

accuracy: 0,957



Implementierung & Evaluation

Briefsammlung – Kategorisierung nach Autorinnen und Autoren

Methode: Bigramme TF-IDF-Matrix

Classifier: LinearSVC-dense

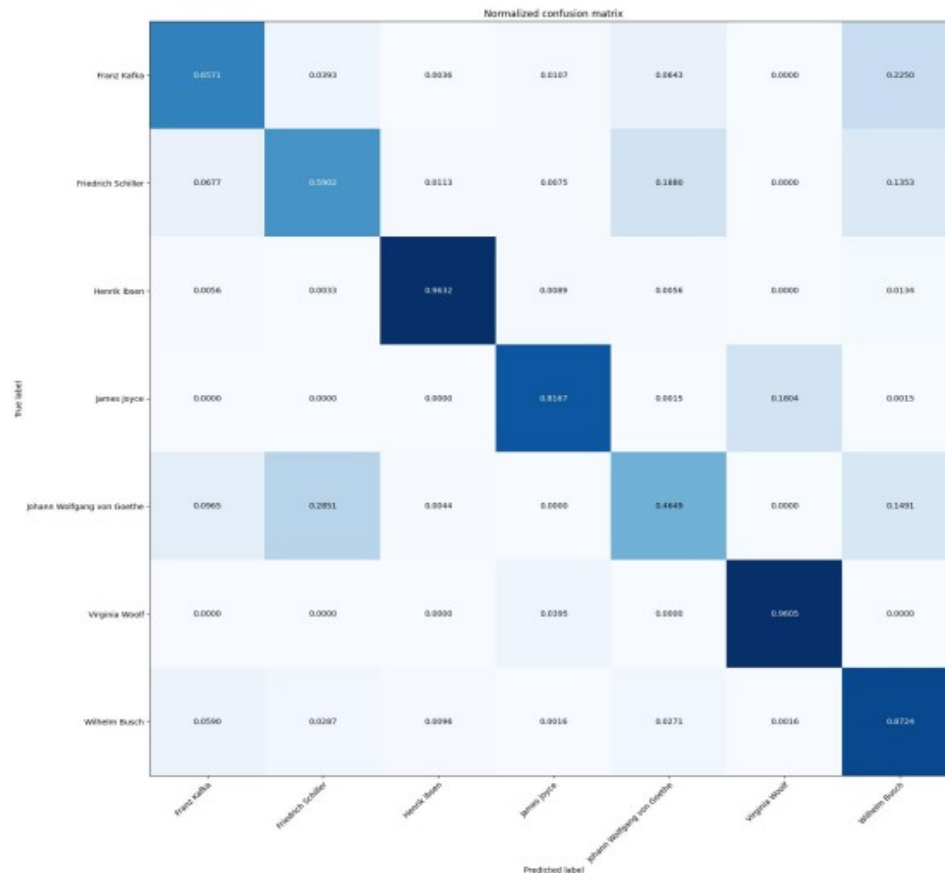
	Precision	recall	f1-score	support
Kafka	0.692	0.657	0.674	280
Schiller	0.618	0.590	0.604	266
Ibsen	0.987	0.963	0.975	897
Joyce	0.862	0.817	0.839	682
Goethe	0.538	0.465	0.499	228
Woolf	0.936	0.961	0.948	1901
Busch	0.789	0.872	0.829	627

accuracy			0.869	4881
macro avg	0.775	0.761	0.767	4881
weighted avg	0.867	0.869	0.867	4881

Ergebnisse

f1-score: Ibsen 0,975, Woolf 0,948, Joyce 0,839

accuracy: 0,869



Implementierung & Evaluation

Briefsammlung – Kategorisierung nach Autorinnen und Autoren

Methode: Bigramme TF-IDF-Matrix

Classifier: LinearSVC-LSA

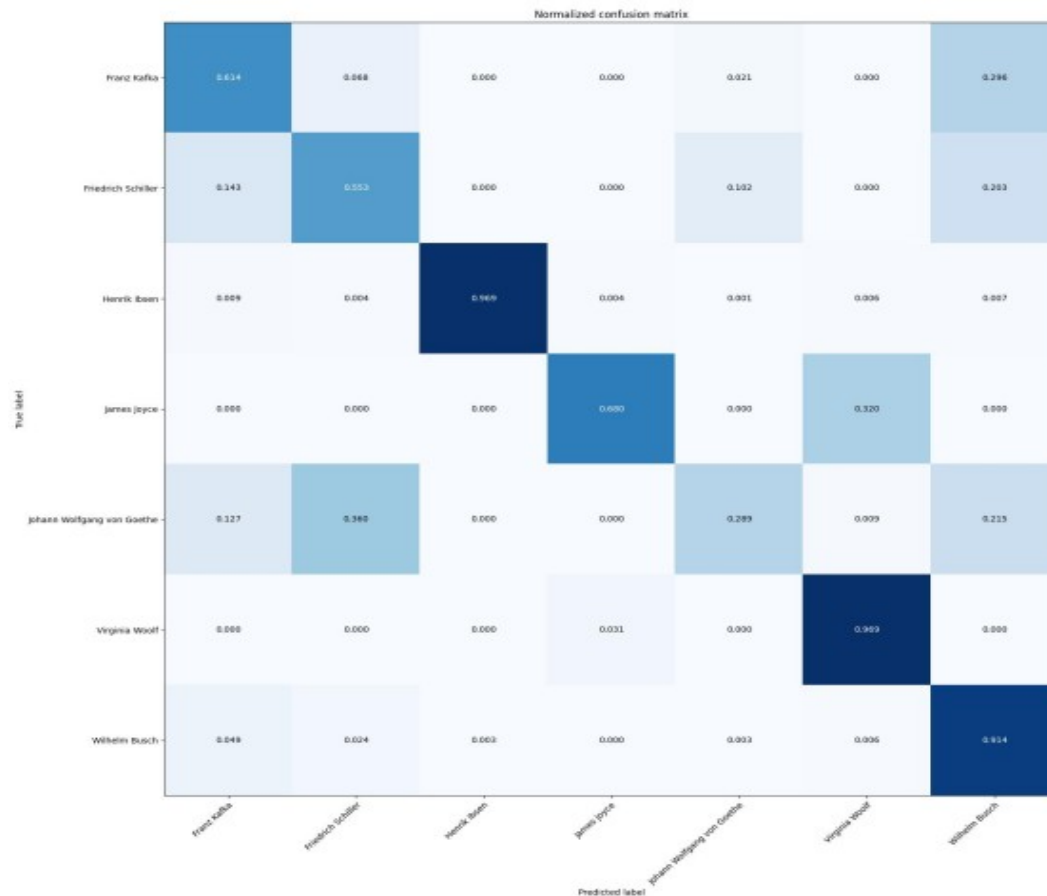
	Precision	recall	f1-score	support
Kafka	0.594	0.611	0.602	280
Schiller	0.529	0.515	0.522	266
Ibsen	0.999	0.969	0.984	897
Joyce	0.892	0.702	0.786	682
Goethe	0.642	0.268	0.378	228
Woolf	0.896	0.972	0.932	1901
Busch	0.738	0.907	0.814	627

accuracy			0.847	4881
macro avg	0.756	0.706	0.717	4881
weighted avg	0.845	0.847	0.839	4881

Ergebnisse

f1-score: Ibsen 0,984, Woolf 0,932, Busch 0,814

accuracy: 0,847



Implementierung & Evaluation

IMDb-Filmrezensionen – Sentiment-Analyse

Methode: wortbasierte TF-IDF-Matrix

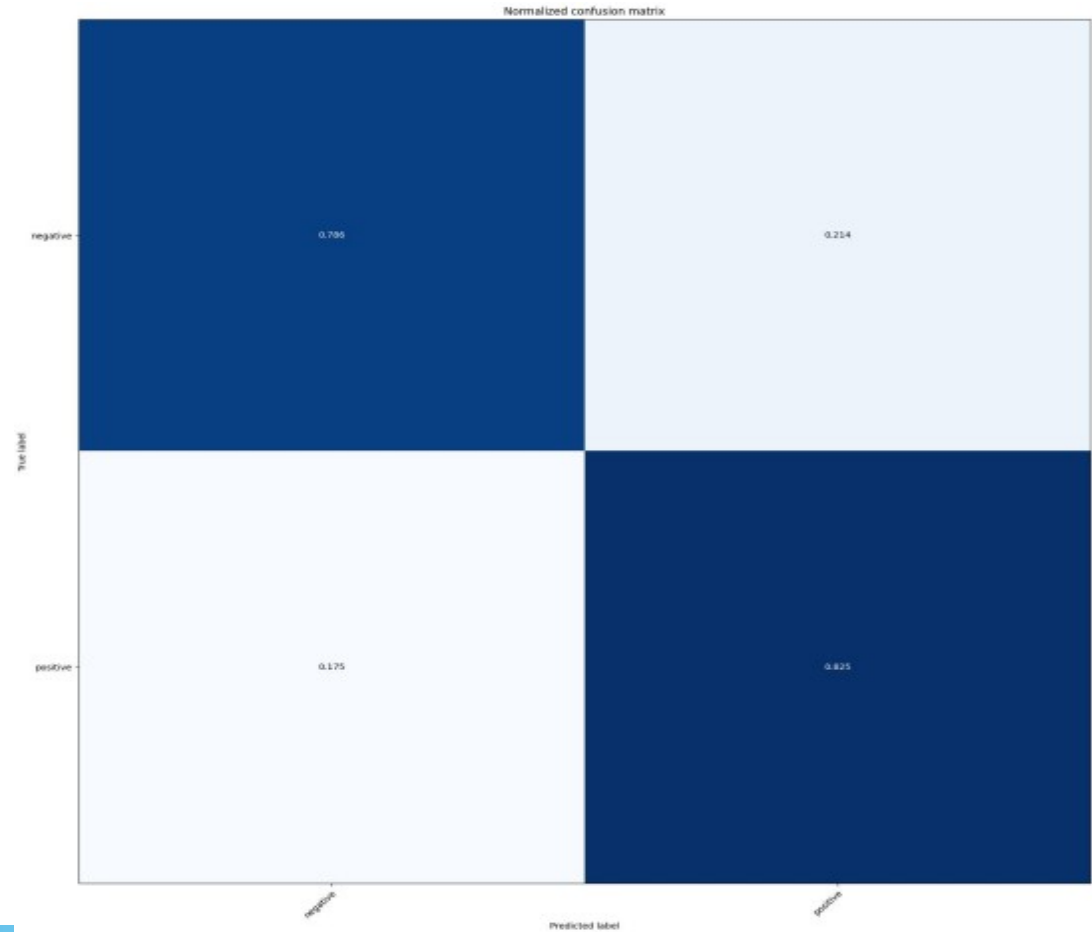
Classifier: LinearSVC-dense

	Precision	Recall	f1-score	Support
negativ	0.817	0.786	0.801	4985
positiv	0.795	0.825	0.810	5015
accuracy			0.806	10000
macro avg	0.806	0.806	0.806	10000
weighted avg	0.806	0.806	0.806	10000

Ergebnisse:

f1-score: Positive 0,810

accuracy: 0,806



Implementierung & Evaluation

IMDb-Filmrezensionen – Sentiment-Analyse

Methode: wortbasierte TF-IDF-Matrix

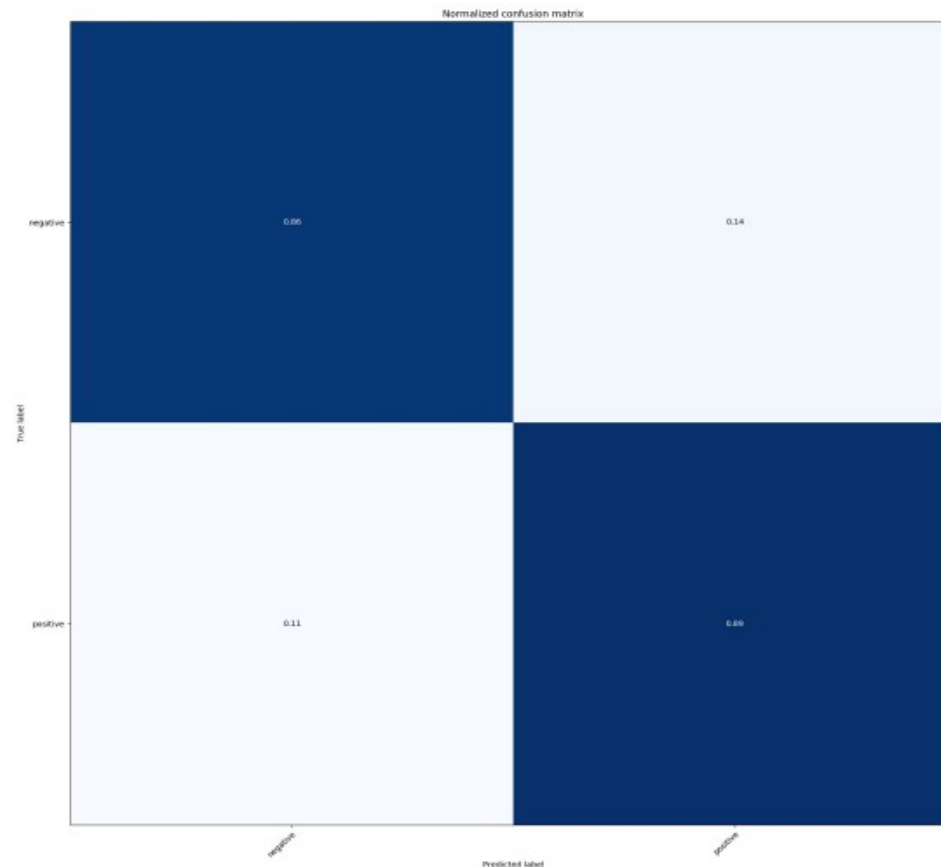
LinearSVC-LSA

	precision	recall	f1-score	support
negative	0.885	0.866	0.875	4985
positive	0.869	0.888	0.878	5015
accuracy			0.877	10000
macro avg	0.877	0.877	0.877	10000
Weighted avg	0.877	0.877	0.877	10000

Ergebnisse:

F1-score: positive 0,878

accuracy: 0,877



Implementierung & Evaluation

IMDb-Filmrezensionen – Sentiment-Analyse

Methode: wortbasierte TF-IDF-Matrix

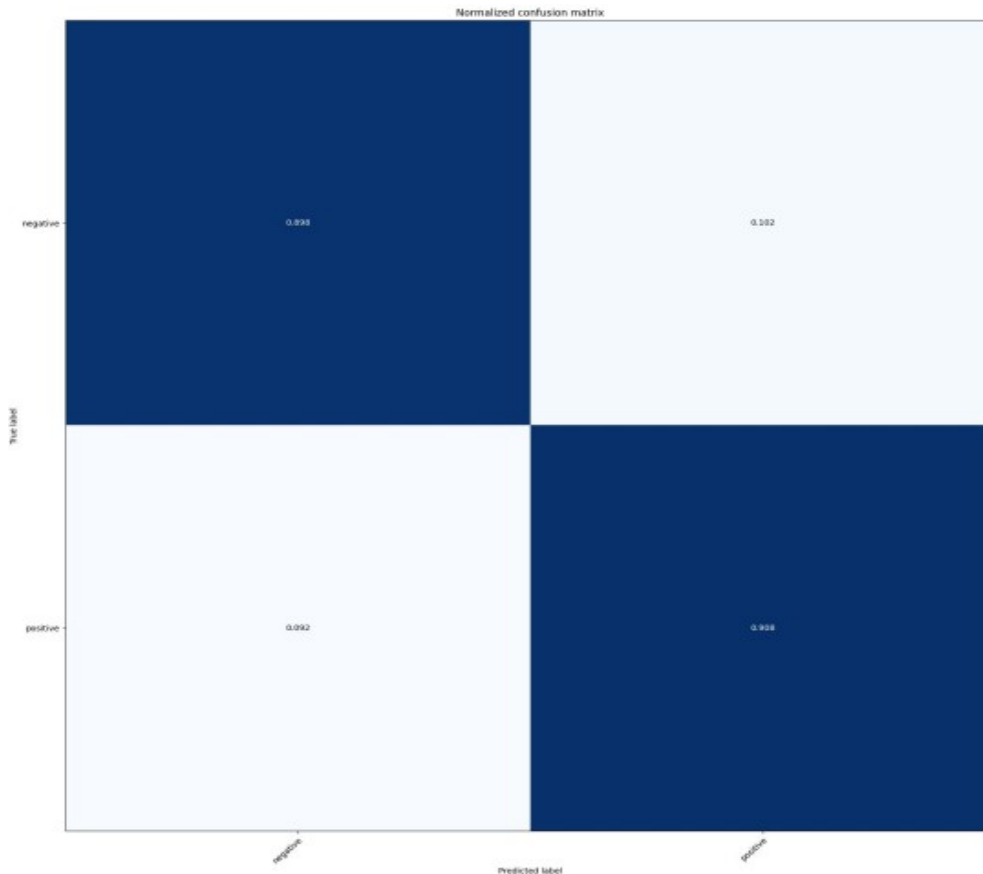
Classifier LinearSVC-sparse

	precision	recall	f1-score	support
negative	0.906	0.898	0.902	4985
positive	0.899	0.908	0.903	5015
accuracy			0.903	10000
macro avg	0.903	0.903	0.903	10000
Weighted avg	0.903	0.903	0.903	10000

Ergebnisse:

F1-score: positive 0,903

accuracy: 0,903



Implementierung & Evaluation

IMDb-Filmrezensionen – Sentiment-Analyse

Methode: Bigramme TF-IDF-Matrix

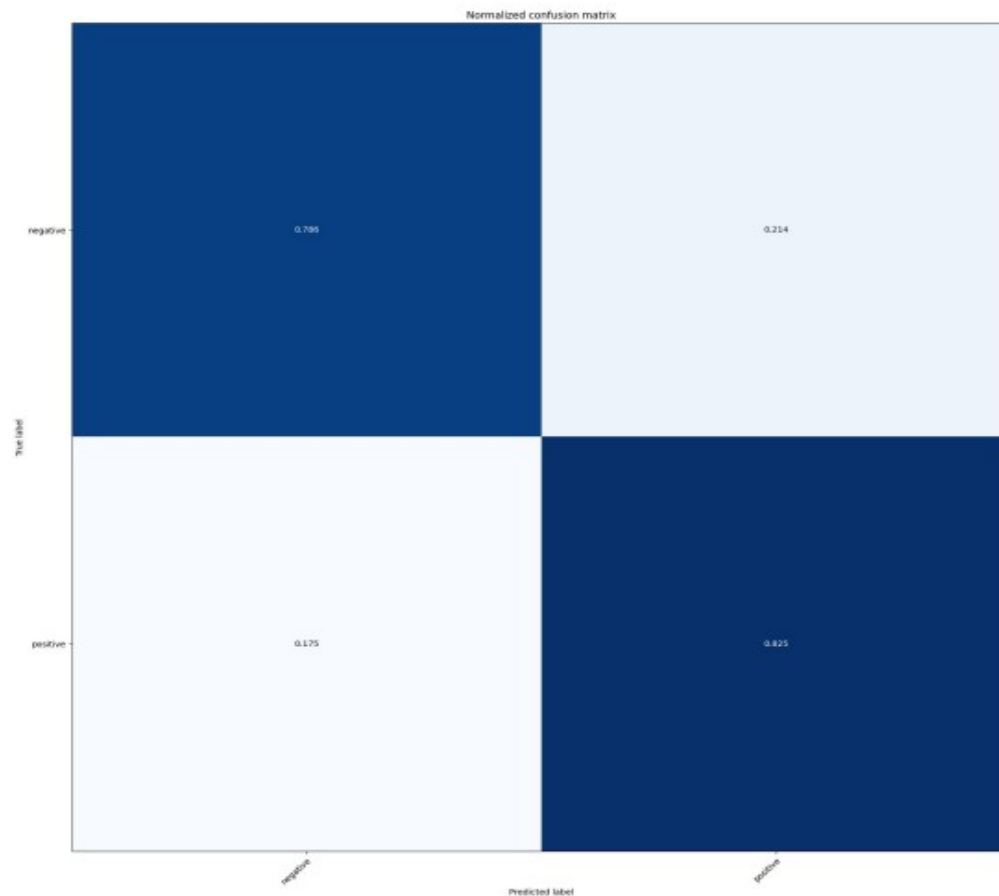
Classifier LinearSVC-dense

	precision	recall	f1-score	support
negative	0.817	0.786	0.801	4985
positive	0.795	0.825	0.810	5015
accuracy			0.806	10000
macro avg	0.806	0.806	0.806	10000
Weighted avg	0.806	0.806	0.806	10000

Ergebnisse

F1-score: positive 0,810

accuracy: 0,806



Implementierung & Evaluation

IMDb-Filmrezensionen – Sentiment-Analyse

Methode: Bigramme TF-IDF-Matrix

Classifier LinearSVC-LSA

	precision	recall	f1-score	support
negative	0.870	0.851	0.860	4985
positive	0.855	0.873	0.864	5015
accuracy			0.862	10000
macro avg	0.862	0.862	0.862	10000
Weighted avg	0.862	0.862	0.862	10000

Ergebnisse

F1-score: positive 0,864

accuracy: 0,862



Implementierung & Evaluation

Huffington Post – Kategorien

Methode: wortbasierte TF-IDF-Matrix

Classifier: LinearSVC-dense

Ergebnisse

f1-score: Weddings 0,764; Divorce 0,733, Style & Beauty 0,714

Accuracy: 0,559

#Classification	report: precision	recall	f1-score	support
ARTS & CULTURE	0.565	0.277	0.371	282
BLACK VOICES	0.481	0.421	0.449	392
BUSINESS	0.509	0.387	0.439	380
COLLEGE	0.574	0.547	0.560	212
COMEDY	0.494	0.409	0.447	399
CRIME	0.568	0.714	0.633	409
CULTURE & ARTS	0.484	0.540	0.510	391
DIVORCE	0.714	0.752	0.733	419
EDUCATION	0.526	0.515	0.520	198
ENTERTAINMENT	0.466	0.491	0.478	387
ENVIRONMENT	0.499	0.580	0.536	355
FIFTY	0.532	0.366	0.434	273
GOOD NEWS	0.518	0.377	0.436	305
HEALTHY LIVING	0.286	0.199	0.235	386
HOME & LIVING	0.667	0.738	0.701	386
IMPACT	0.439	0.295	0.353	400
MEDIA	0.559	0.633	0.594	395
MONEY	0.551	0.645	0.595	324
PARENTING	0.473	0.542	0.505	391
POLITICS	0.499	0.505	0.502	420
QUEER VOICES	0.717	0.660	0.688	415
RELIGION	0.601	0.736	0.662	440
SCIENCE	0.607	0.643	0.624	414
SPORTS	0.610	0.778	0.684	410
STYLE	0.549	0.540	0.545	413
STYLE & BEAUTY	0.623	0.627	0.625	391
TASTE	0.639	0.809	0.714	397
TECH	0.572	0.690	0.625	416
TRAVEL	0.594	0.691	0.639	405
WEDDINGS	0.734	0.795	0.764	400
WEIRD NEWS	0.481	0.338	0.397	408
WELLNESS	0.411	0.356	0.382	407
WOMEN	0.484	0.445	0.464	400
WORLD	0.645	0.710	0.676	404
accuracy			0.559	12824
macro avg	0.549	0.552	0.545	12824
weighted avg	0.551	0.559	0.550	12824

Implementierung & Evaluation

Huffington Post – Kategorien

Methode: wortbasierte TF-IDF-Matrix

Classifier: LinearSVC-LSA

Ergebnisse

f1-score: Weddings 0,653, Divorce 0,560, Crime 0,472

Accuracy: 0,317

#Classification	report: precision	recall	f1-score	support
ARTS & CULTURE	0.250	0.004	0.007	282
BLACK VOICES	0.448	0.207	0.283	392
BUSINESS	0.251	0.166	0.200	380
COLLEGE	0.468	0.344	0.397	212
COMEDY	0.302	0.263	0.281	399
CRIME	0.361	0.680	0.472	409
CULTURE & ARTS	0.201	0.322	0.248	391
DIVORCE	0.658	0.487	0.560	419
EDUCATION	0.292	0.247	0.268	198
ENTERTAINMENT	0.160	0.106	0.127	387
ENVIRONMENT	0.296	0.192	0.232	355
FIFTY	0.208	0.110	0.144	273
GOOD NEWS	0.233	0.115	0.154	305
HEALTHY LIVING	0.185	0.145	0.163	386
HOME & LIVING	0.408	0.567	0.475	386
IMPACT	0.123	0.045	0.066	400
MEDIA	0.325	0.319	0.322	395
MONEY	0.267	0.312	0.288	324
PARENTING	0.303	0.432	0.357	391
POLITICS	0.291	0.340	0.314	420
QUEER VOICES	0.502	0.352	0.414	415
RELIGION	0.333	0.248	0.284	440
SCIENCE	0.301	0.411	0.348	414
SPORTS	0.241	0.256	0.248	410
STYLE	0.280	0.354	0.313	413
STYLE & BEAUTY	0.376	0.588	0.459	391
TASTE	0.309	0.496	0.381	397
TECH	0.352	0.418	0.382	416
TRAVEL	0.286	0.259	0.272	405
WEDDINGS	0.628	0.680	0.653	400
WEIRD NEWS	0.202	0.147	0.170	408
WELLNESS	0.208	0.182	0.194	407
WOMEN	0.306	0.333	0.319	400
WORLD	0.231	0.379	0.287	404
accuracy			0.317	12824
macro avg	0.311	0.309	0.296	12824
weighted avg	0.313	0.317	0.302	12824

Implementierung & Evaluation

Huffington Post – Kategorien

Methode: wortbasierte TF-IDF-Matrix

Classifier: LinearSVC-sparse

Ergebnisse

f1-score: Weddings 0,757, Culture & Arts 0,736, Home & Living 0,718

Accuracy: 0,526

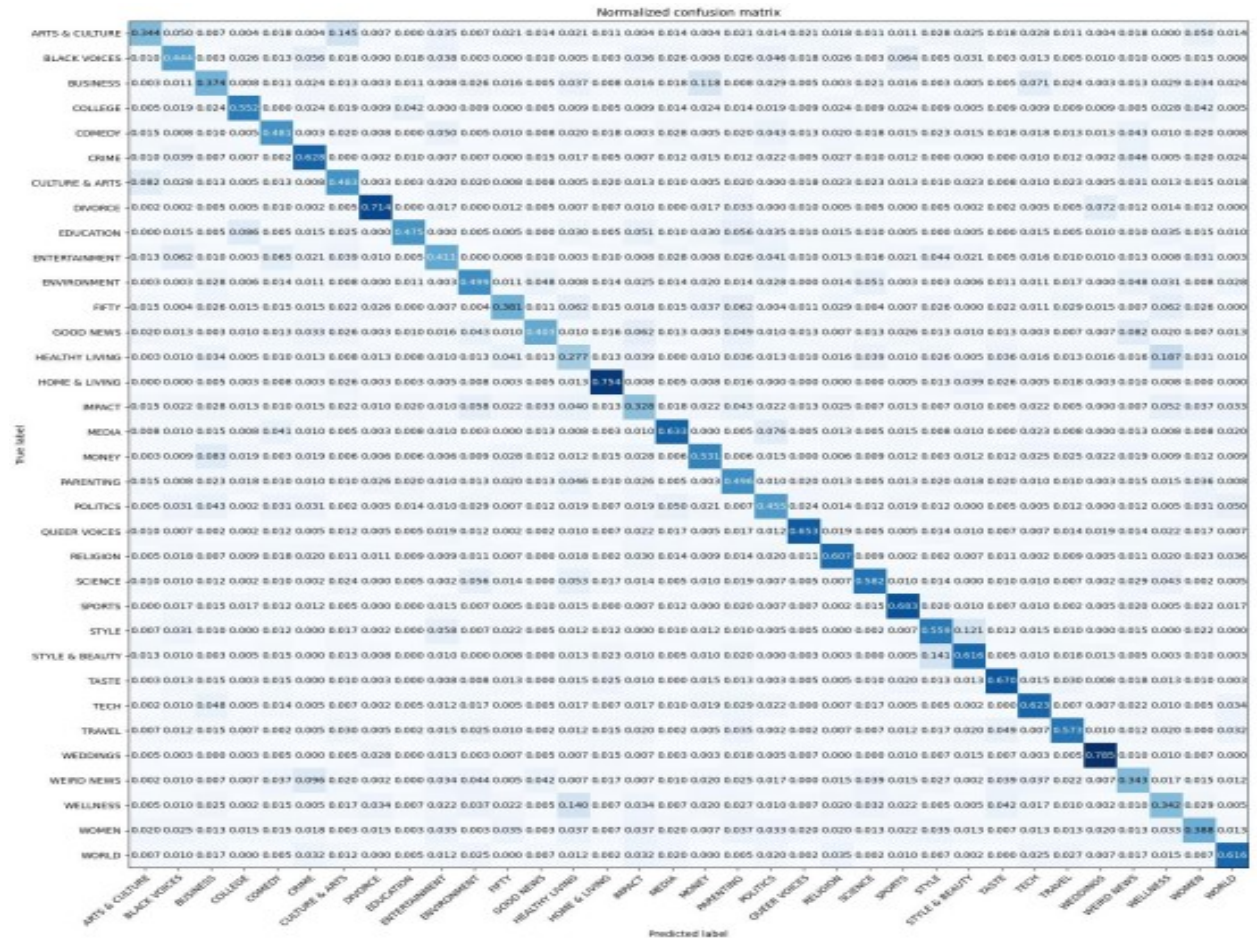
#Classification	report: precision	recall	f1-score	support
ARTS & CULTURE	0.441	0.344	0.386	282
BLACK VOICES	0.468	0.444	0.455	392
BUSINESS	0.406	0.374	0.389	380
COLLEGE	0.522	0.552	0.537	212
COMEDY	0.509	0.481	0.495	399
CRIME	0.572	0.628	0.599	409
CULTURE & ARTS	0.472	0.483	0.478	391
DIVORCE	0.761	0.714	0.736	419
EDUCATION	0.534	0.475	0.503	198
ENTERTAINMENT	0.436	0.411	0.423	387
ENVIRONMENT	0.471	0.499	0.484	355
FIFTY	0.419	0.381	0.399	273
GOOD NEWS	0.500	0.403	0.446	305
HEALTHY LIVING	0.277	0.277	0.277	386
HOME & LIVING	0.685	0.754	0.718	386
IMPACT	0.360	0.328	0.343	400
MEDIA	0.611	0.633	0.622	395
MONEY	0.490	0.531	0.510	324
PARENTING	0.413	0.496	0.451	391
POLITICS	0.465	0.455	0.460	420
QUEER VOICES	0.715	0.653	0.683	415
RELIGION	0.618	0.607	0.612	440
SCIENCE	0.603	0.582	0.592	414
SPORTS	0.625	0.683	0.653	410
STYLE	0.518	0.559	0.538	413
STYLE & BEAUTY	0.575	0.616	0.595	391
TASTE	0.627	0.670	0.648	397
TECH	0.583	0.623	0.602	416
TRAVEL	0.587	0.573	0.580	405
WEDDINGS	0.730	0.785	0.757	400
WEIRD NEWS	0.374	0.343	0.358	408
WELLNESS	0.335	0.342	0.338	407
WOMEN	0.406	0.388	0.396	400
WORLD	0.587	0.616	0.601	404
accuracy			0.526	12824
macro avg	0.520	0.521	0.520	12824
weighted avg	0.524	0.526	0.524	12824

Implementierung & Evaluation

Huffington Post – Kategorien

Methode: wortbasierte TF-IDF-Matrix

Classifier: LinearSVC-sparse



Implementierung & Evaluation

Huffington Post – Kategorien

Methode: Bigramme TF-IDF-Matrix

Classifier: LinearSVC-dense

Ergebnisse:

f1-score: Weddings 0,764, Culture & Arts 0,733, Taste 0,714

Accuracy: 0,559

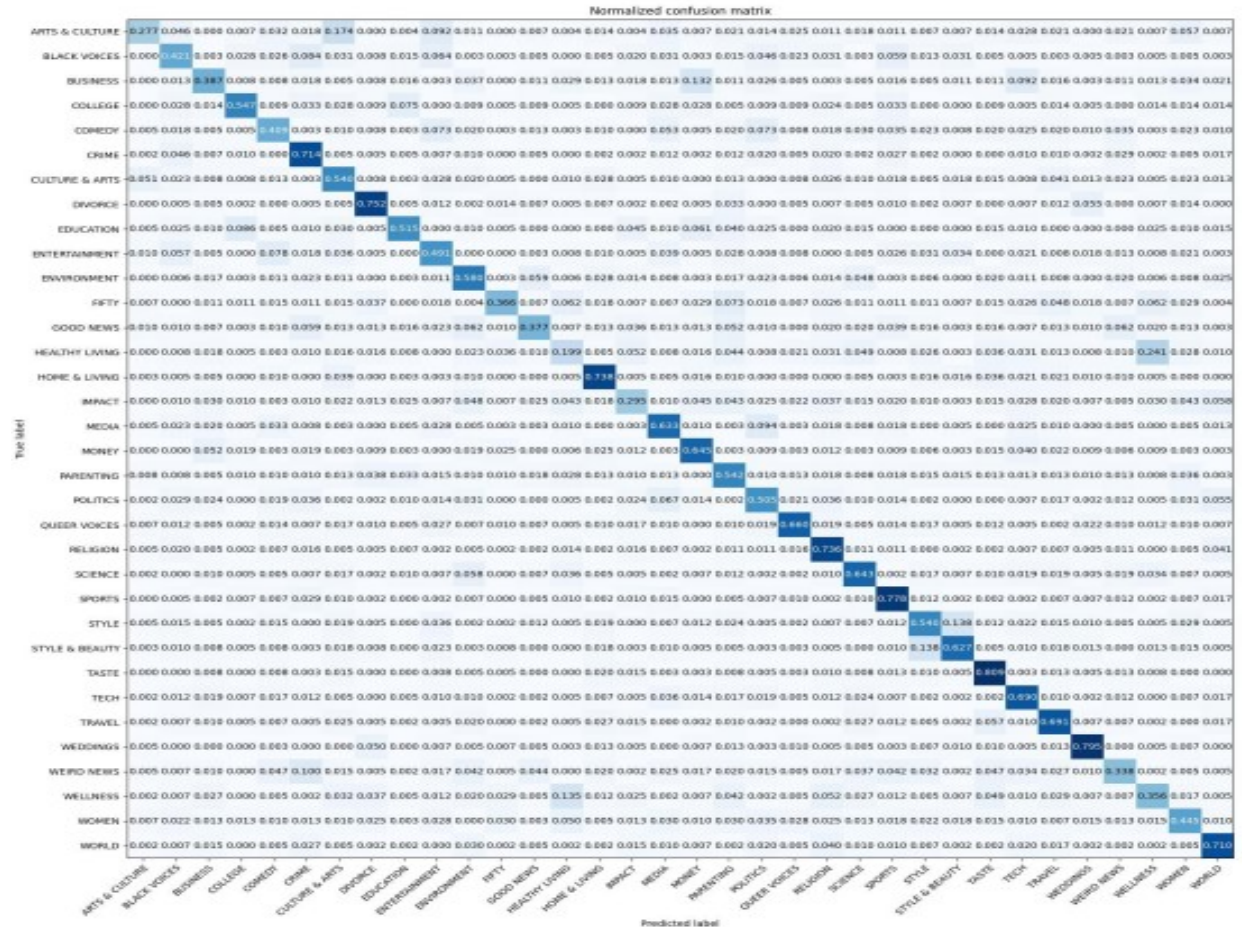
#Classification	report: precision	recall	f1-score	support
ARTS & CULTURE	0.565	0.277	0.371	282
BLACK VOICES	0.481	0.421	0.449	392
BUSINESS	0.509	0.387	0.439	380
COLLEGE	0.574	0.547	0.560	212
COMEDY	0.494	0.409	0.447	399
CRIME	0.568	0.714	0.633	409
CULTURE & ARTS	0.484	0.540	0.510	391
DIVORCE	0.714	0.752	0.733	419
EDUCATION	0.526	0.515	0.520	198
ENTERTAINMENT	0.466	0.491	0.478	387
ENVIRONMENT	0.499	0.580	0.536	355
FIFTY	0.532	0.366	0.434	273
GOOD NEWS	0.518	0.377	0.436	305
HEALTHY LIVING	0.286	0.199	0.235	386
HOME & LIVING	0.667	0.738	0.701	386
IMPACT	0.439	0.295	0.353	400
MEDIA	0.559	0.633	0.594	395
MONEY	0.551	0.645	0.595	324
PARENTING	0.473	0.542	0.505	391
POLITICS	0.499	0.505	0.502	420
QUEER VOICES	0.717	0.660	0.688	415
RELIGION	0.601	0.736	0.662	440
SCIENCE	0.607	0.643	0.624	414
SPORTS	0.610	0.778	0.684	410
STYLE	0.549	0.540	0.545	413
STYLE & BEAUTY	0.623	0.627	0.625	391
TASTE	0.639	0.809	0.714	397
TECH	0.572	0.690	0.625	416
TRAVEL	0.594	0.691	0.639	405
WEDDINGS	0.734	0.795	0.764	400
WEIRD NEWS	0.481	0.338	0.397	408
WELLNESS	0.411	0.356	0.382	407
WOMEN	0.484	0.445	0.464	400
WORLD	0.645	0.710	0.676	404
accuracy			0.559	12824
macro avg	0.549	0.552	0.545	12824
weighted avg	0.551	0.559	0.550	12824

Implementierung & Evaluation

Huffington Post – Kategorien

Methode: Bigramme TF-IDF-Matrix

Classifier: LinearSVC-dense



Implementierung & Evaluation

Huffington Post – Kategorien

Methode: Bigramme TF-IDF-Matrix

Classifier: LinearSVC-LSA

Ergebnisse:

f1-score: Home & Living 0,328, Style & Beauty 0,278, Divorce 0,235

Accuracy: 0,157

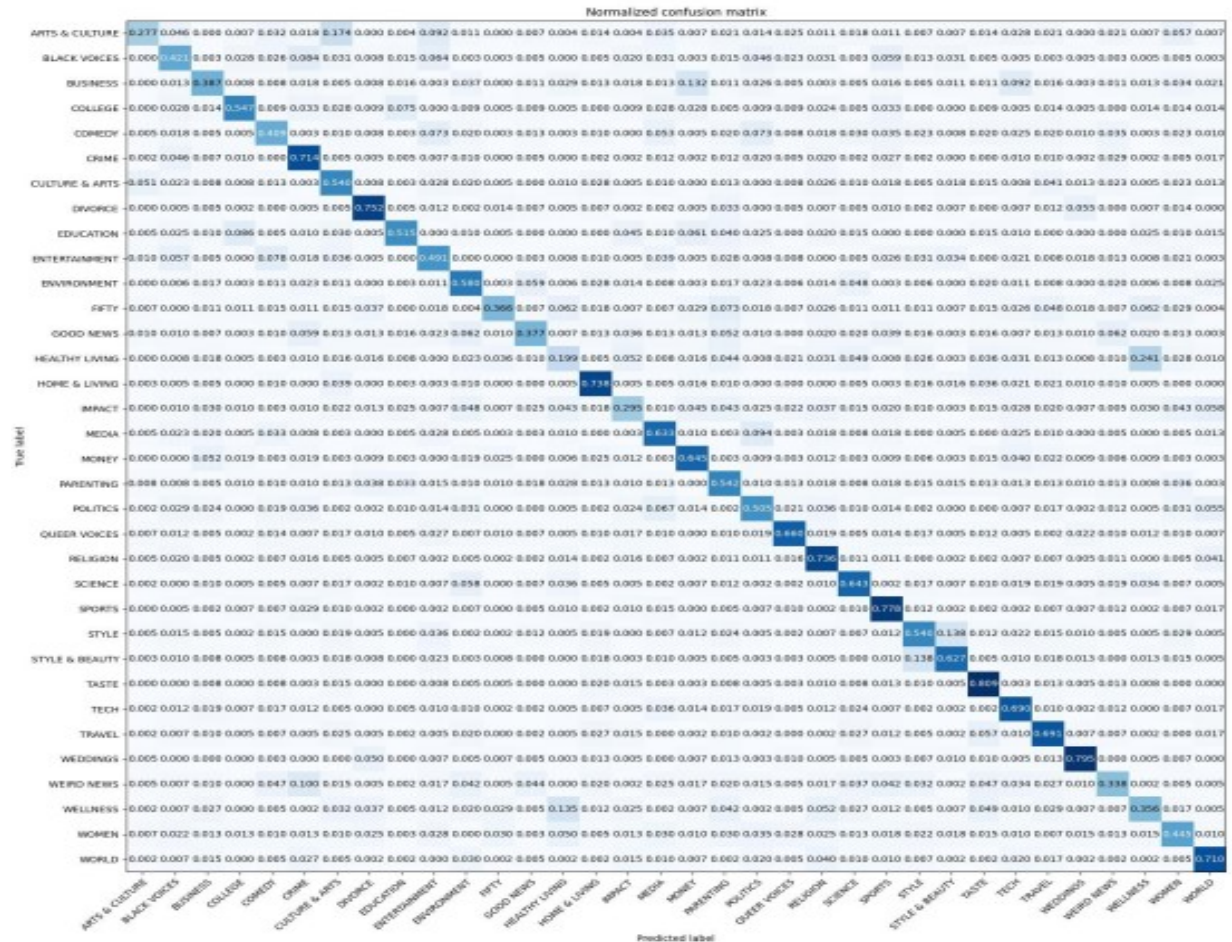
#Classification	report: precision	recall	f1-score	support
ARTS & CULTURE	0.083	0.011	0.019	282
BLACK VOICES	0.178	0.074	0.105	392
BUSINESS	0.278	0.118	0.166	380
COLLEGE	0.486	0.085	0.145	212
COMEDY	0.198	0.243	0.218	399
CRIME	0.082	0.685	0.147	409
CULTURE & ARTS	0.156	0.184	0.169	391
DIVORCE	0.233	0.236	0.235	419
EDUCATION	0.200	0.035	0.060	198
ENTERTAINMENT	0.107	0.059	0.077	387
ENVIRONMENT	0.245	0.144	0.181	355
FIFTY	0.089	0.015	0.025	273
GOOD NEWS	0.056	0.003	0.006	305
HEALTHY LIVING	0.101	0.101	0.101	386
HOME & LIVING	0.274	0.407	0.328	386
IMPACT	0.109	0.045	0.064	400
MEDIA	0.220	0.218	0.219	395
MONEY	0.164	0.071	0.099	324
PARENTING	0.184	0.100	0.129	391
POLITICS	0.198	0.181	0.189	420
QUEER VOICES	0.182	0.089	0.120	415
RELIGION	0.306	0.150	0.201	440
SCIENCE	0.192	0.205	0.199	414
SPORTS	0.115	0.100	0.107	410
STYLE	0.154	0.140	0.147	413
STYLE & BEAUTY	0.259	0.299	0.278	391
TASTE	0.222	0.232	0.227	397
TECH	0.194	0.087	0.120	416
TRAVEL	0.125	0.128	0.127	405
WEDDINGS	0.229	0.188	0.206	400
WEIRD NEWS	0.081	0.064	0.071	408
WELLNESS	0.141	0.066	0.090	407
WOMEN	0.205	0.105	0.139	400
WORLD	0.128	0.240	0.167	404
accuracy			0.157	12824
macro avg	0.182	0.150	0.143	12824
weighted avg	0.180	0.157	0.148	12824

Implementierung & Evaluation

Huffington Post – Kategorien

Methode: Bigramme TF-IDF-Matrix

Classifier: LinearSVC-LSA



Zusammenfassung der Ergebnisse

Methode	Wortbasierte tf-idf		
	dense	LSA	sparse
Sprachidentifizierung			
f1-score	de 0.999, en 0.999, da 0.996	de 0.999, en 0.998, da 0.994	en 0.999, de 0.998, da 0.996
accuracy	accuracy: 0.997	accuracy: 0.995	accuracy: 0.996
Autoren			
f1-score	Ibsen 0.975, Woolf 0.948, Joyce 0.839	Ibsen 0.985, Woolf 0.959, Busch 0.890	Ibsen 0.993, Woolf 0.982, Busch 0.965
accuracy	accuracy: 0.869	accuracy: 0.898	accuracy: 0.957
Kategorien			
f1-score	Weddings 0.764, Divorce 0.733, Style& Beauty 0.714	Weddings 0.653, Divorce 0.560, Crime 0.472	Weddings 0.757, Culture & Arts 0.736, Home & Living 0.718
accuracy	accuracy: 0.559	accuracy: 0.317	accuracy: 0.526
Sentiment Analyse			
f1-score	positive 0.810	positive 0.878	positive 0.903
accuracy	accuracy: 0.806	accuracy: 0.877	accuracy: 0.903

Zusammenfassung der Ergebnisse

Methode	Bigramme	
	dense	LSA
Sprachidentifizierung f1-score	en 0.999, de 0.999, da 0.996	en 0.997, de 0.992, da 0.994
accuracy	accuracy: 0.997	accuracy: 0.990
Autoren f1-score	Ibsen 0.975, Woolf 0.948, Joyce 0.839	Ibsen 0.984, Woolf 0.932, Busch 0.814
accuracy	accuracy: 0.869	accuracy: 0.847
Kategorien f1-score	Weddings 0.764, Culture & Arts 0.733, Taste 0.714	Home & Living 0.328, Style & Beauty 0.278, Divorce 0.235
accuracy	accuracy: 0.559	accuracy: 0.157
Sentiment Analyse f1-score	positive 0.810	positive 0.864
accuracy	accuracy: 0.806	accuracy: 0.862

Literatur

Bernhard Boser, Isabelle Guyon, Vladimir Vapnik, A training algorithm for optimal margin classifiers, COLT '92: Proceedings of the fifth annual workshop on Computational learning theory, 1992, S. 144–152.

Cortes, Vapnik, Support-vector networks, Machine Learning, Band 20, 1995, S. 273–297.

Nello Cristianini, John Shawe-Taylor: Kernel Methods for Pattern Analysis, Cambridge University Press, Cambridge, 2004.

R.A. Fisher, THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS, Annals of Eugenics, vol. 7, nr. 2, p. 179-188, (1936) : Trennungskonzept, Iris-Datensatz [<https://onlinelibrary.wiley.com/doi/epdf/10.1111/j.1469-1809.1936.tb02137.x>].

Manning, C. and Schütze, H. 2001. Foundations of Statistical Natural Language Processing. MIT Press, Cambridge MA.

Rosenblatt, F. (1958): The Perceptron, a Probabilistic Model for Information Storage and Organisation in the Brain, in Psychological Review, 62/386, S. 386–408.

Ingo Steinwart, Andreas Christmann: Support Vector Machines, Springer, New York, 2008.

Vapnik und Tschervonenkis: Theorie der Mustererkennung, 1979.

Vielen Dank für Ihre Aufmerksamkeit!