



Support vector machine

Group G:

Huihui Hu

Liuxu Lu

Bingyu Xiong

Dozent:

Stefan Langer



Support vector machine

1. Hintergrund von SVM
2. Was ist SVM Algorithmus ? Wie arbeit SVM ?
3. Wie kann SVM im Python angewandte werden ?
4. Vor und Nachteile, Entwicklung von SVM

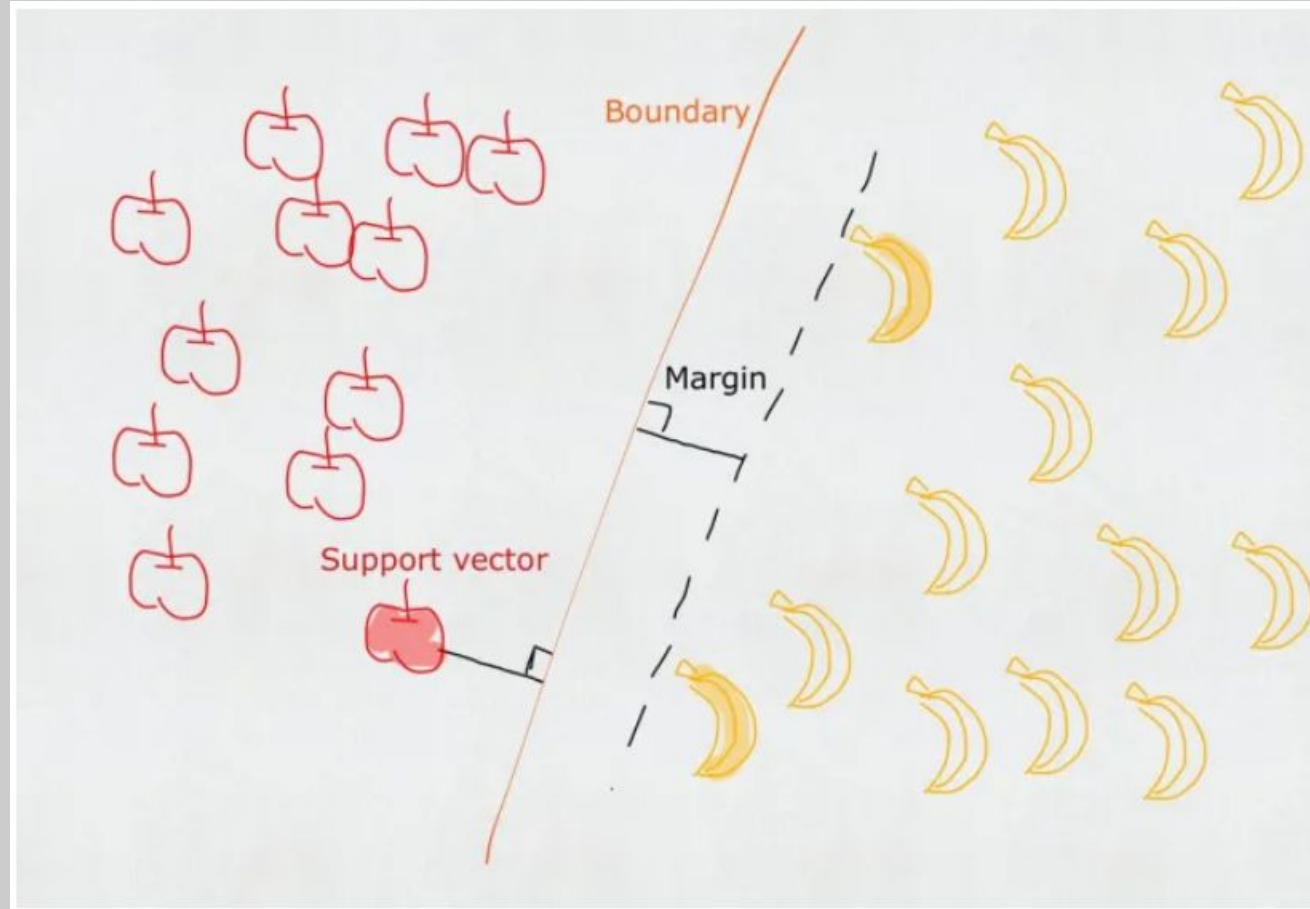
Überblick

Es wurde erstmals 1963 von Vladimir N. Vapnik und Alexey Ya. Chervanenkis vorgeschlagen.

Die aktuelle Version von Soft Margin wurde 1993 von Corinna Cortes und Vapnik vorgeschlagen und 1995 veröffentlicht.

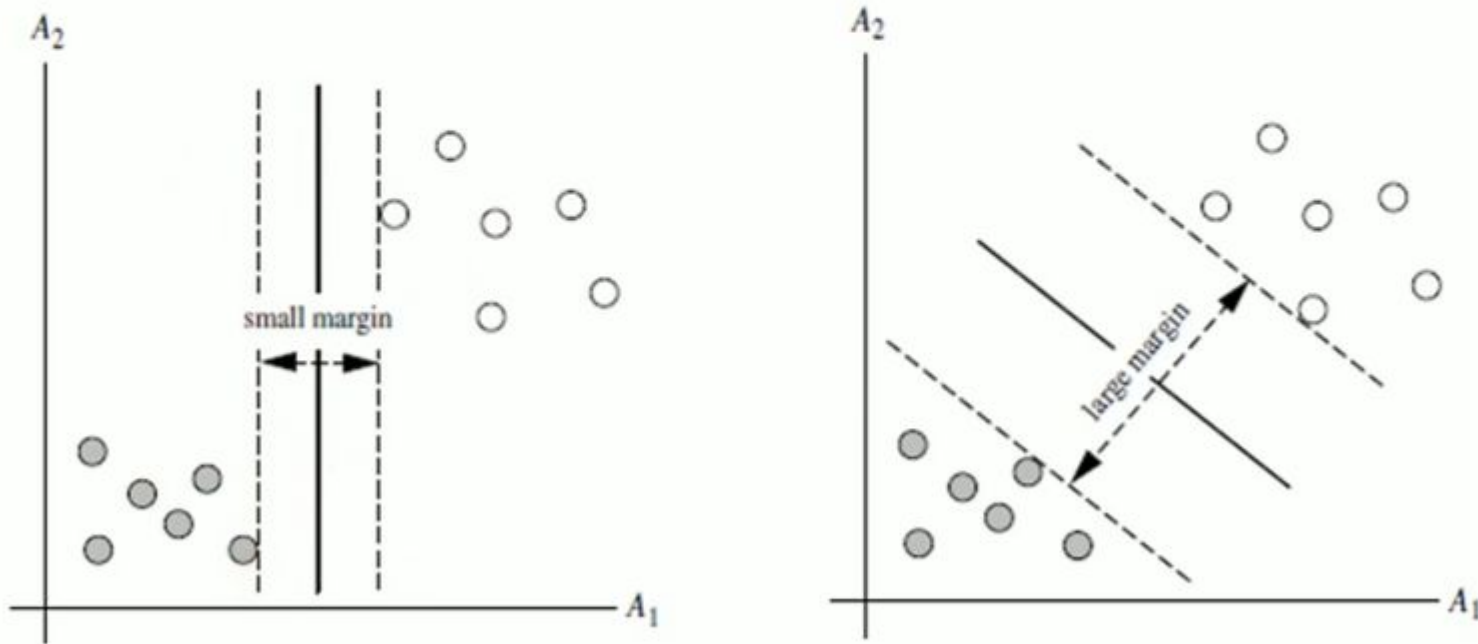
Vor dem Aufkommen von Deep Learning im Jahr 2012 galt svm als der erfolgreichste und leistungsfähigste Algorithmus im maschinellen Lernen der letzten zehn Jahre

Was ist Support Vector?

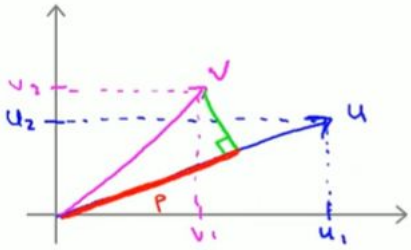


Mathematische Grundlagen

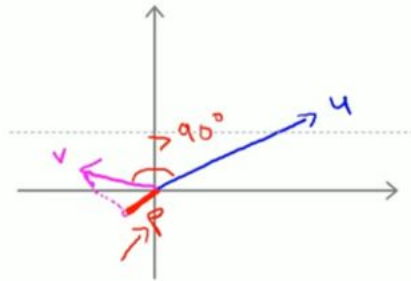
SVM sucht nach einer Hyperebene, die die beiden Klassen unterscheidet und den margin maximiert



Mathematische Grundlagen



Skalarprodukt: $v \cdot u > 0$



Skalarprodukt : $v \cdot u < 0$

$$x = \begin{Bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{Bmatrix} \quad y = \begin{Bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{Bmatrix}$$

Skalarprodukt: $x \cdot y = x_1 y_1 + x_2 y_2 + \dots + x_n y_n$

Skalarprodukt: $x \cdot y = \|x\| \|y\| \cos(\theta)$

Norm : $\|x\| = \sqrt{x \cdot x} = \sqrt{x_1^2 + x_2^2 + \dots + x_n^2}$

$$\cos \theta = \frac{x \cdot y}{\|x\| \|y\|}$$

Mathematische Grundlagen

$$\begin{aligned}w \cdot x_1 + b &= 1 \\w \cdot x_2 + b &= -1\end{aligned}$$

$$\begin{aligned}w \cdot (x_1 - x_2) &= 2 \\ \|w\| \|x_1 - x_2\| \cos(\theta) &= 2 \\ \|w\| \cdot d &= 2 \\ d &= \frac{2}{\|w\|}\end{aligned}$$

$$\begin{aligned}w \cdot x + b &\geq 1, \quad y = 1 \\ w \cdot x + b &\leq -1, \quad y = -1\end{aligned} \quad \left. \vphantom{\begin{aligned}w \cdot x + b &\geq 1, \quad y = 1 \\ w \cdot x + b &\leq -1, \quad y = -1\end{aligned}} \right\} y(w \cdot x + b) \geq 1$$

Gesucht: $\min \frac{\|w\|^2}{2}$

Mathematische Grundlagen

Zielfunktion Lagrange-Multiplikatoren Beschränkungen

$$L(w, b, a) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^n \alpha_i (y_i (w^T x_i + b) - 1)$$


$$\frac{\partial L}{\partial w} = 0 \rightarrow w = \sum_{i=1}^n \alpha_i y_i x_i$$

$$\frac{\partial L}{\partial b} = 0 \rightarrow \sum_{i=1}^n \alpha_i y_i = 0$$

Mathematische Grundlagen

$$\frac{\partial L}{\partial w} = 0 \rightarrow w = \sum_{i=1}^n \alpha_i y_i x_i$$

$$\frac{\partial L}{\partial b} = 0 \rightarrow \sum_{i=1}^n \alpha_i y_i = 0$$

Eliminiere w und b

把w和b消除了

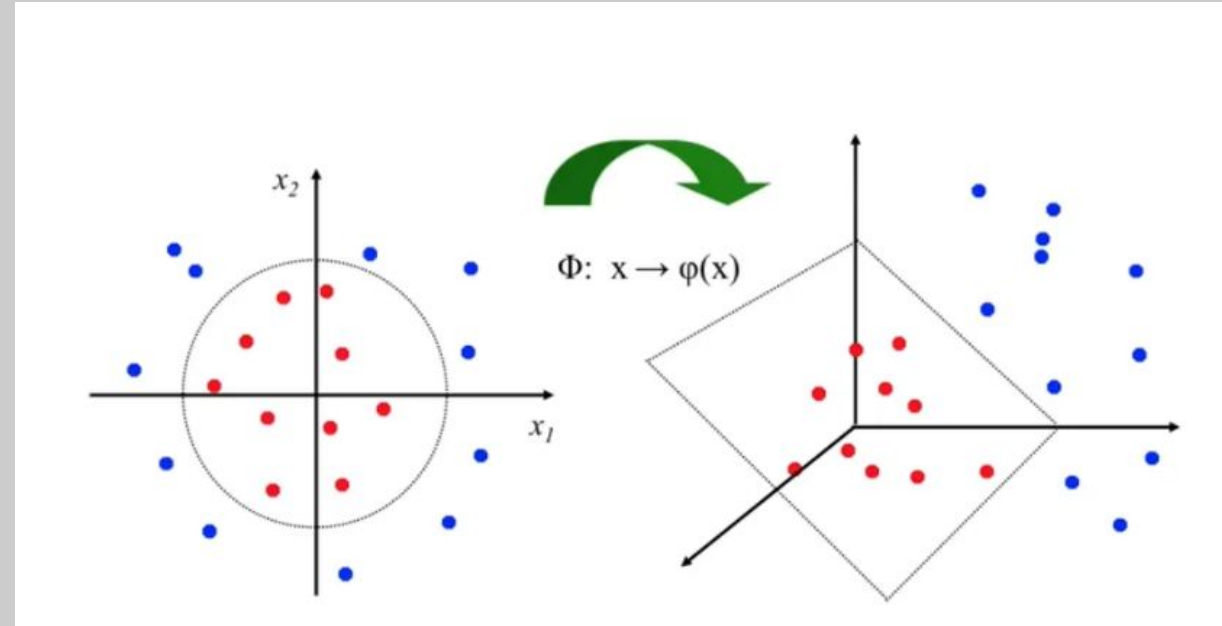
→ $L(w, b, a) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j x_i^T x_j$

$$w^* = \sum_{i=1}^n \alpha_i^* y_i x_i$$

$$b^* = y_i - (w^*)^T x_i$$

Was ist Kernel-Funktion?

Die Kernel-Technik bedeutet, dass durch die Einführung einer Kernel-Funktion das n-dimensionale niedrigdimensionale nichtlineare Problem in ein m-dimensionales hoch-dimensionales lineares Problem umgewandelt werden kann, um es zu lösen. Es muss nur n-mal multipliziert werden, was viele Berechnungen reduziert.



Was ist Schlupfvariable?

Schlupfvariable ξ

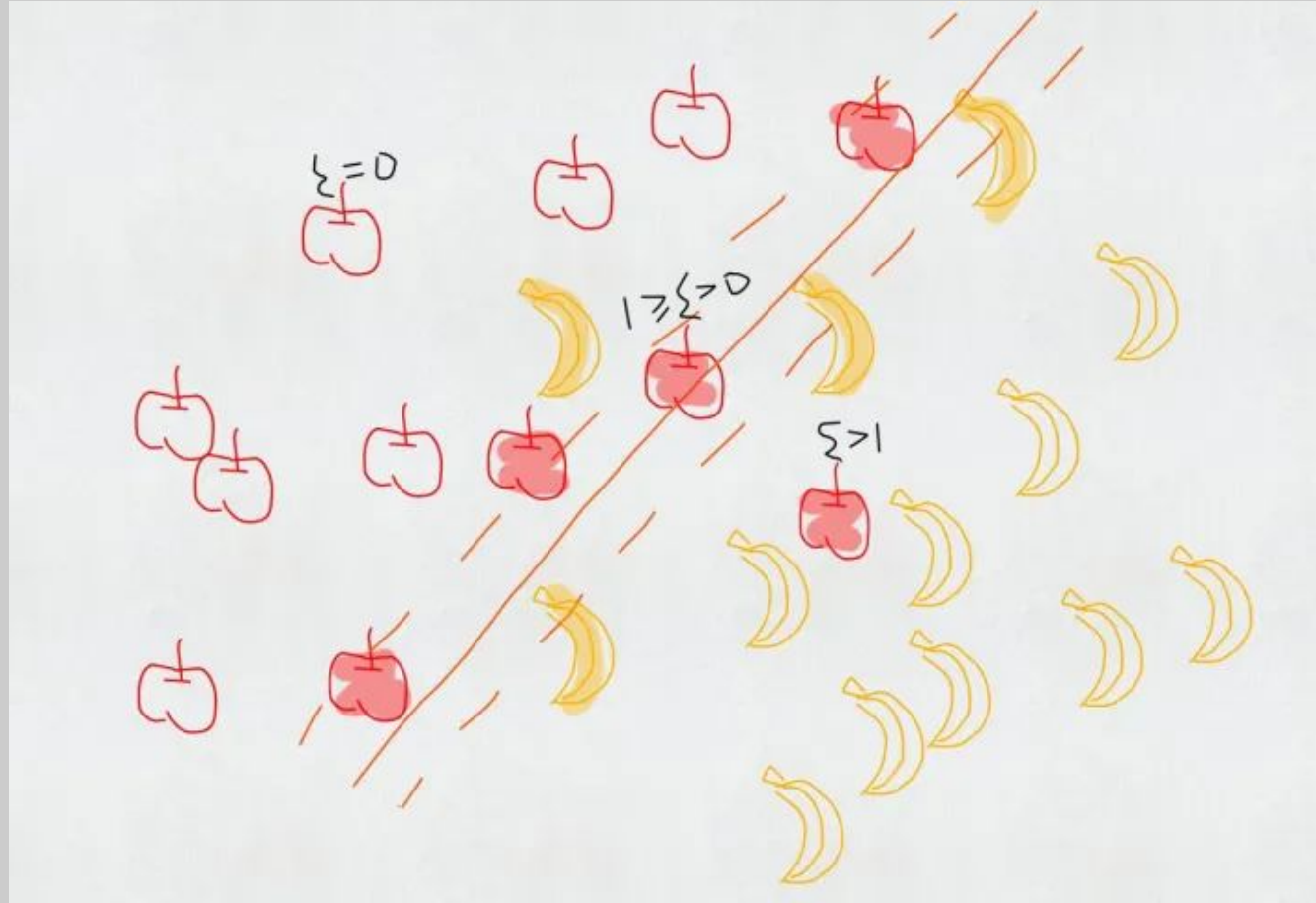
Innerhalb von Margin:

richtig eingeordnete Apfel ($0 < \xi < 1$)

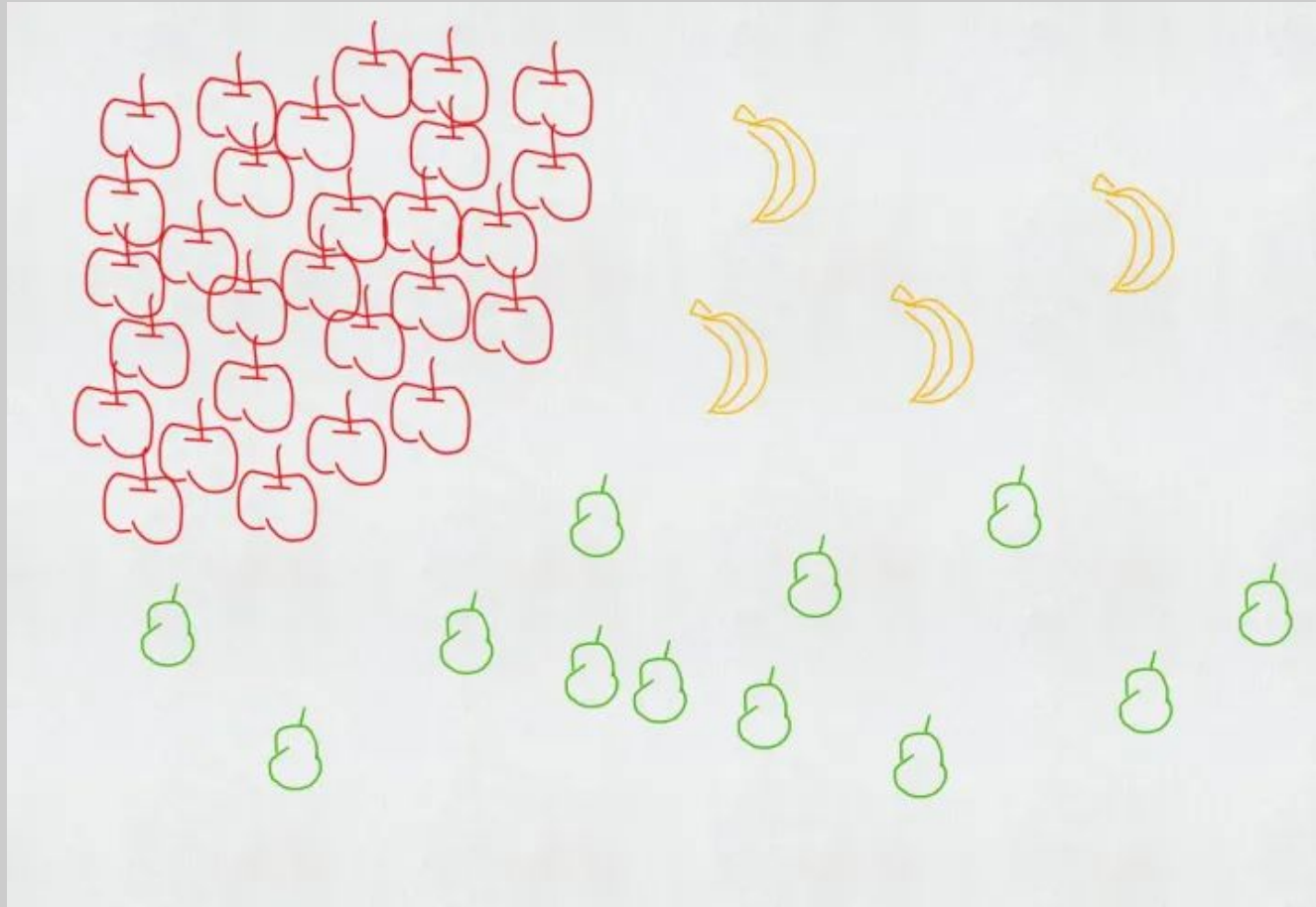
Außerhalb von Margin:

falsch eingeordnete Apfel ($\xi > 1$)

richtig eingeordnet Apfel ($\xi = 1$)



SVM bei Multi-Classification



Apfel, Nicht Apfel
Banana, Nicht Banana
Birne, Nicht Birne
.....

author classification on letter dataset

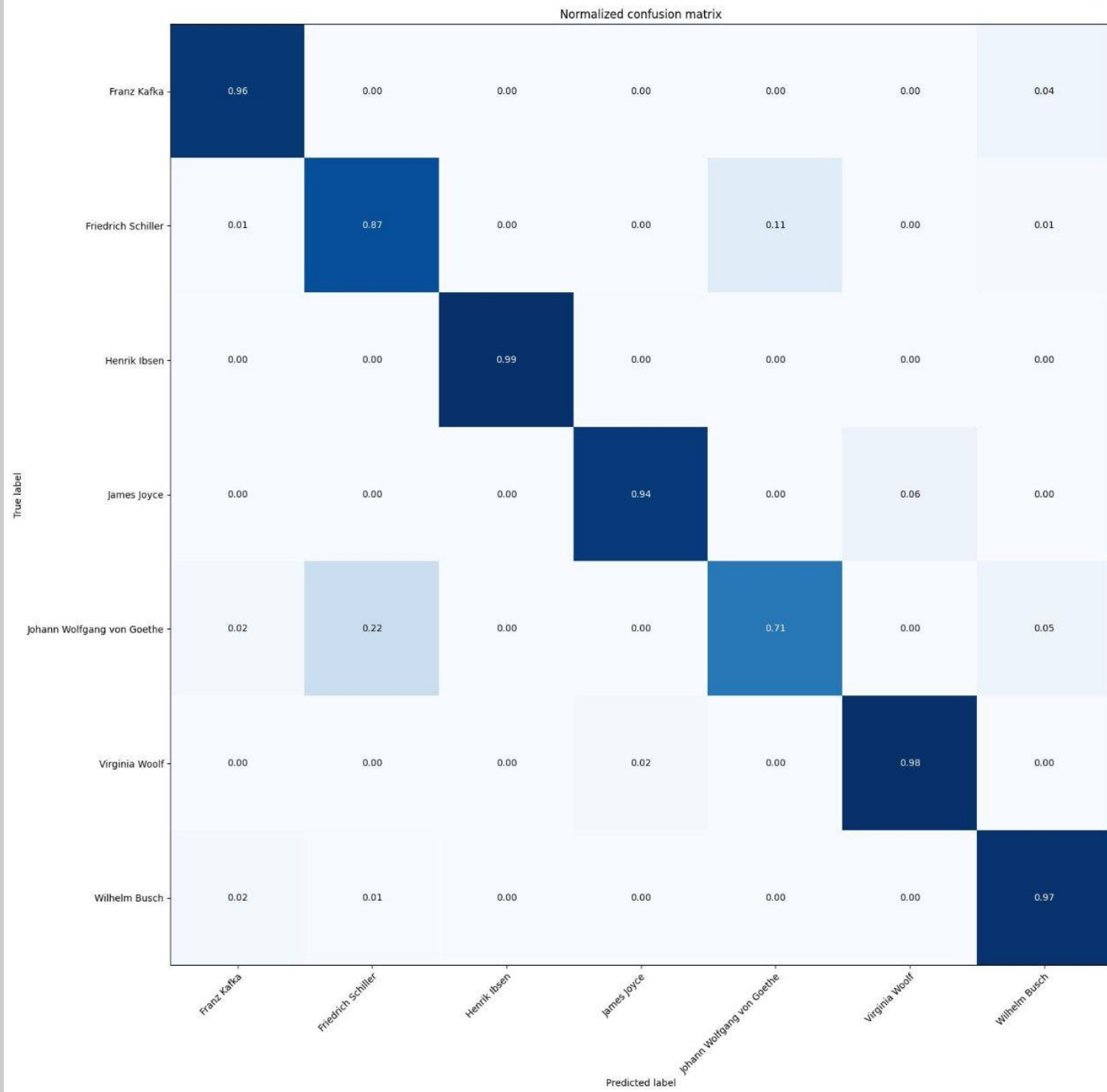
```
#Info:
Classifier: LinearSVC
Label: author

#Counts:
Number of training data_records: 39077
Number of classified data_records: 4881
Number of unique classes in data_records: 7
Number of unique classes found: 7

#Performance:
Seconds used for training: 119
Seconds used for classification: 9

#Classification report:
```

	precision	recall	f1-score	support
Franz Kafka	0.93	0.96	0.95	280
Friedrich Schiller	0.80	0.87	0.84	266
Henrik Ibsen	1.00	0.99	0.99	897
James Joyce	0.95	0.94	0.95	682
Johann Wolfgang von Goethe	0.82	0.71	0.76	228
Virginia Woolf	0.98	0.98	0.98	1901
Wilhelm Busch	0.96	0.97	0.97	627
accuracy			0.96	4881
macro avg	0.92	0.92	0.92	4881
weighted avg	0.96	0.96	0.96	4881



language classification on letter dataset

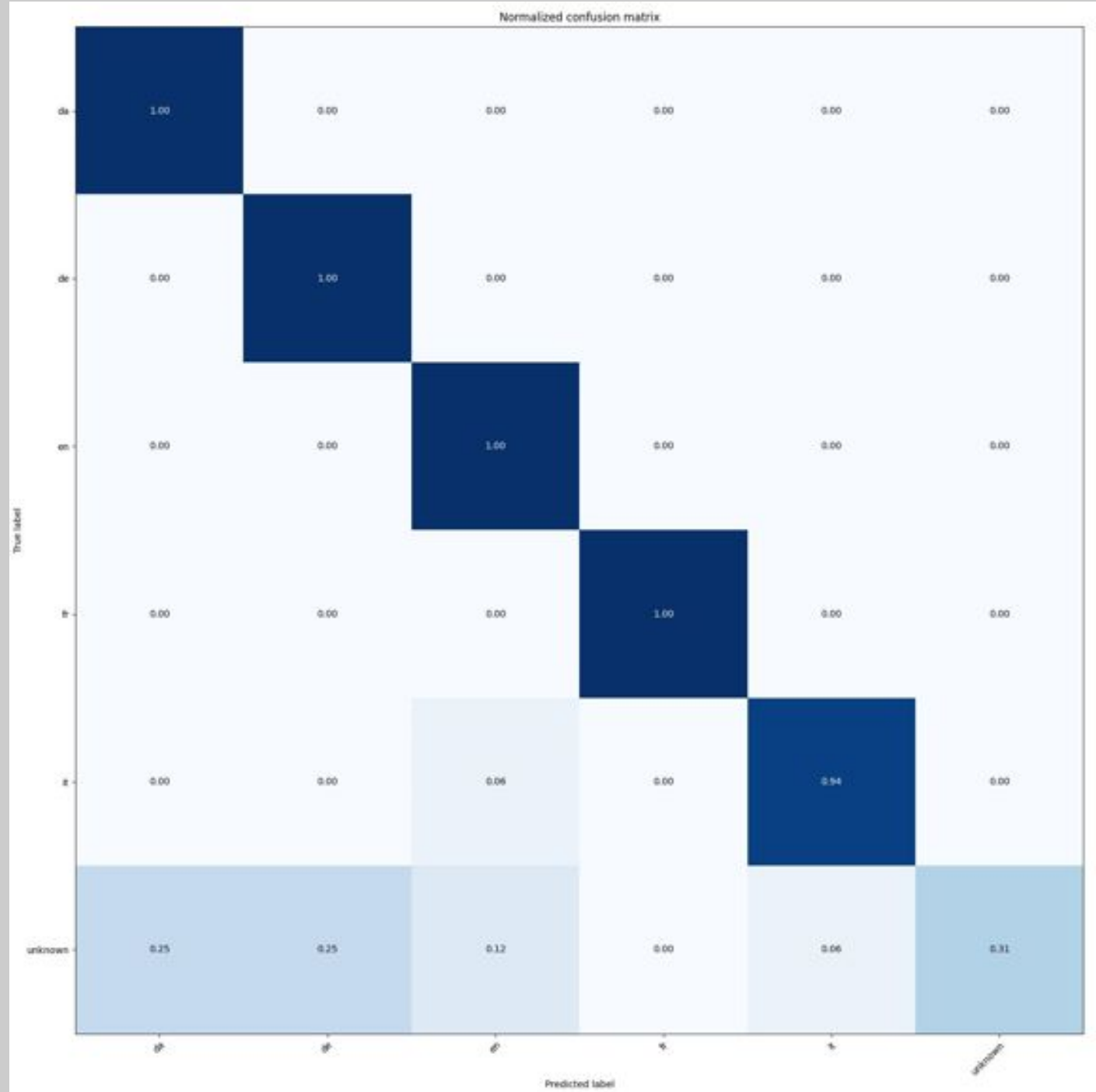
```
#Info:
Classifier: LinearSVC
Label: lang

#Counts:
Number of training data_records: 39077
Number of classified data_records: 4881
Number of unique classes in data_records: 6
Number of unique classes found: 6

#Performance:
Seconds used for training: 106
Seconds used for classification: 9

#Classification report:
```

	precision	recall	f1-score	support
da	0.99	1.00	1.00	838
de	1.00	1.00	1.00	1443
en	1.00	1.00	1.00	2517
fr	0.95	1.00	0.97	36
it	0.94	0.94	0.94	31
unknown	0.83	0.31	0.45	16
accuracy			1.00	4881
macro avg	0.95	0.87	0.89	4881
weighted avg	1.00	1.00	1.00	4881



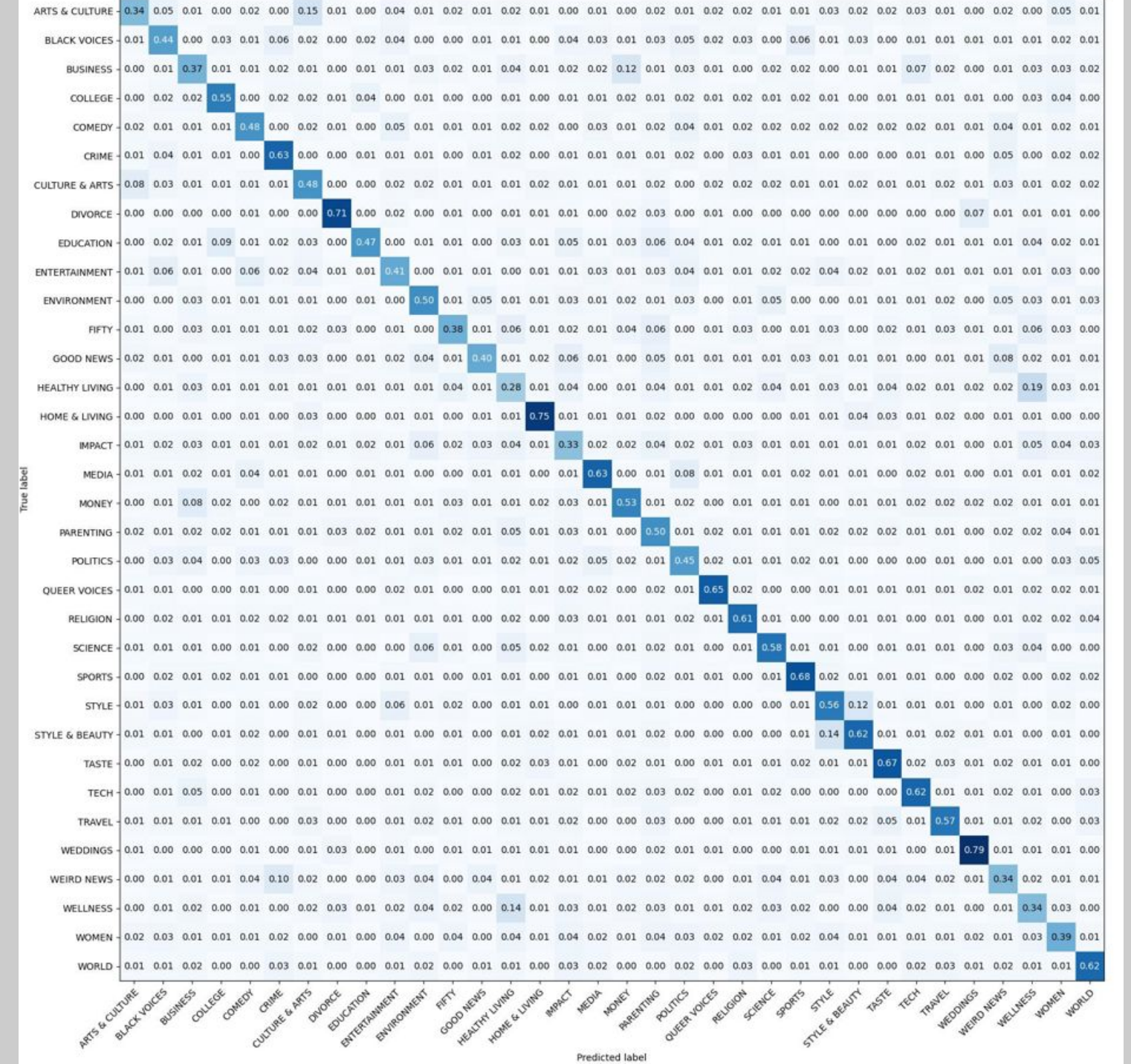
classification on news dataset based on headline

```
#Performance:
Seconds used for training: 52
Seconds used for classification: 28

#Classification report:
              precision    recall  f1-score   support

ARTS & CULTURE      0.44      0.34      0.39       282
  BLACK VOICES      0.47      0.44      0.46       392
    BUSINESS      0.41      0.37      0.39       380
      COLLEGE      0.52      0.55      0.54       212
      COMEDY      0.51      0.48      0.49       399
      CRIME      0.57      0.63      0.60       409
CULTURE & ARTS      0.47      0.48      0.48       391
  DIVORCE      0.76      0.71      0.74       419
  EDUCATION      0.53      0.47      0.50       198
ENTERTAINMENT      0.44      0.41      0.42       387
  ENVIRONMENT      0.47      0.50      0.48       355
    FIFTY      0.42      0.38      0.40       273
    GOOD NEWS      0.50      0.40      0.45       305
HEALTHY LIVING      0.28      0.28      0.28       386
  HOME & LIVING      0.68      0.75      0.72       386
    IMPACT      0.36      0.33      0.34       400
    MEDIA      0.61      0.63      0.62       395
    MONEY      0.49      0.53      0.51       324
  PARENTING      0.41      0.50      0.45       391
  POLITICS      0.46      0.45      0.46       420
QUEER VOICES      0.72      0.65      0.68       415
  RELIGION      0.62      0.61      0.61       440
  SCIENCE      0.60      0.58      0.59       414
  SPORTS      0.62      0.68      0.65       410
  STYLE      0.52      0.56      0.54       413
STYLE & BEAUTY      0.58      0.62      0.60       391
  TASTE      0.63      0.67      0.65       397
  TECH      0.58      0.62      0.60       416
  TRAVEL      0.59      0.57      0.58       405
  WEDDINGS      0.73      0.79      0.76       400
  WEIRD NEWS      0.37      0.34      0.36       408
  WELLNESS      0.33      0.34      0.34       407
  WOMEN      0.41      0.39      0.40       400
  WORLD      0.59      0.62      0.60       404

accuracy              0.53      12824
macro avg              0.52      12824
weighted avg           0.52      12824
```



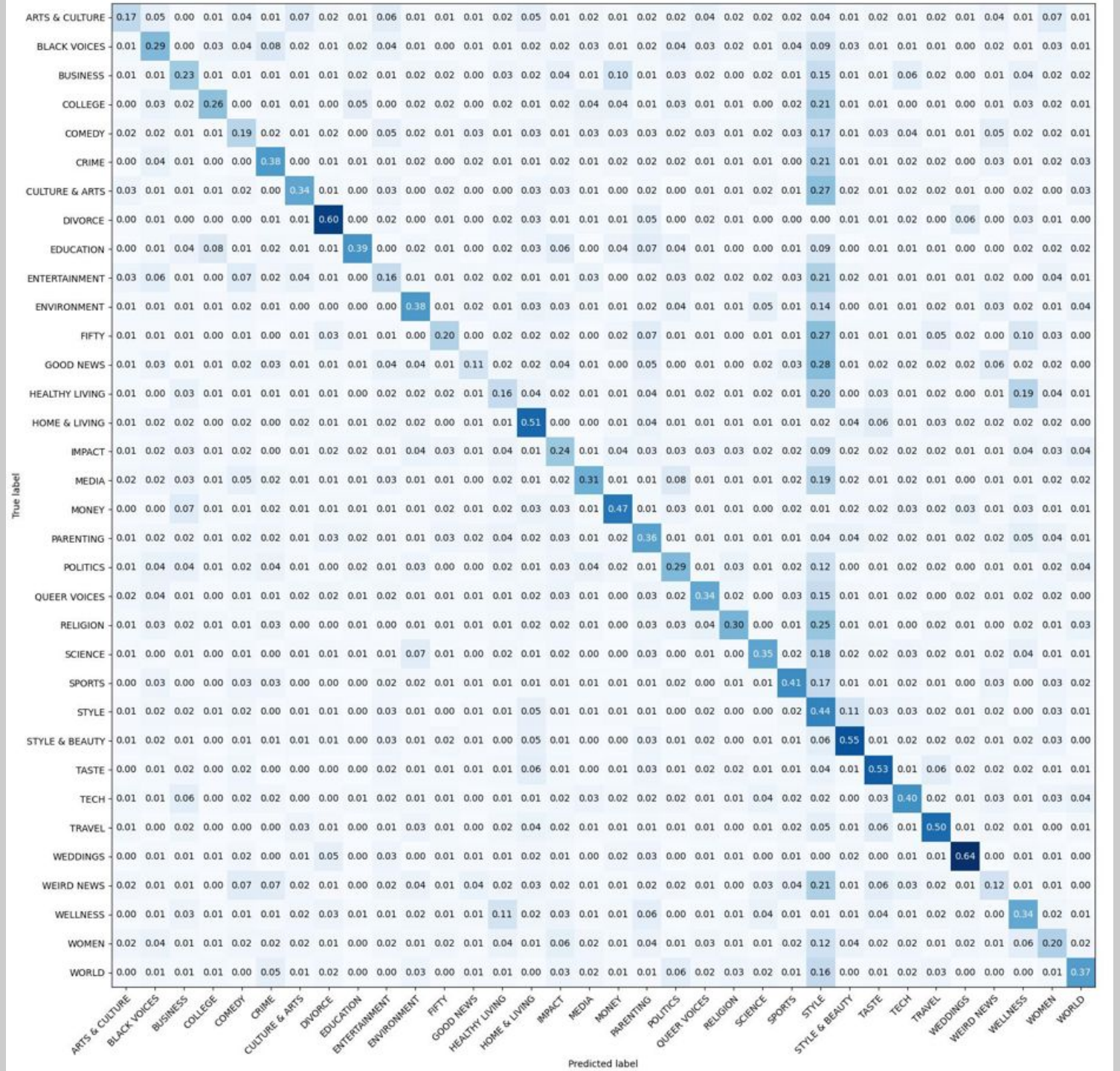
category classification on news dataset based on description

```
#Performance:
Seconds used for training: 68
Seconds used for classification: 28

#Classification report:
              precision    recall  f1-score   support

ARTS & CULTURE      0.27      0.17      0.21       282
  BLACK VOICES      0.31      0.29      0.30       392
    BUSINESS      0.27      0.23      0.25       380
    COLLEGE      0.35      0.26      0.30       212
    COMEDY      0.23      0.19      0.21       399
    CRIME      0.42      0.38      0.40       409
CULTURE & ARTS      0.44      0.34      0.39       391
  DIVORCE      0.62      0.60      0.61       419
  EDUCATION      0.44      0.39      0.41       198
ENTERTAINMENT      0.20      0.16      0.18       387
  ENVIRONMENT      0.39      0.38      0.38       355
    FIFTY      0.28      0.20      0.23       273
  GOOD NEWS      0.22      0.11      0.14       305
HEALTHY LIVING      0.20      0.16      0.17       386
  HOME & LIVING      0.41      0.51      0.45       386
    IMPACT      0.26      0.24      0.25       400
    MEDIA      0.43      0.31      0.36       395
    MONEY      0.44      0.47      0.45       324
  PARENTING      0.30      0.36      0.33       391
  POLITICS      0.33      0.29      0.31       420
  QUEER VOICES      0.41      0.34      0.37       415
  RELIGION      0.48      0.30      0.37       440
  SCIENCE      0.44      0.35      0.39       414
  SPORTS      0.44      0.41      0.43       410
  STYLE      0.10      0.44      0.17       413
STYLE & BEAUTY      0.50      0.55      0.52       391
  TASTE      0.48      0.53      0.50       397
  TECH      0.42      0.40      0.41       416
  TRAVEL      0.47      0.50      0.48       405
  WEDDINGS      0.63      0.64      0.64       400
  WEIRD NEWS      0.18      0.12      0.14       408
  WELLNESS      0.29      0.34      0.31       407
  WOMEN      0.24      0.20      0.22       400
  WORLD      0.44      0.37      0.40       404

accuracy              0.34      12824
macro avg              0.36      0.34      0.34      12824
weighted avg           0.36      0.34      0.35      12824
```

sentiment classification

#Info:

Classifier: LinearSVC

Label: sentiment

#Counts:

Number of training data_records: 40000

Number of classified data_records: 10000

Number of unique classes in data_records: 2

Number of unique classes found: 2

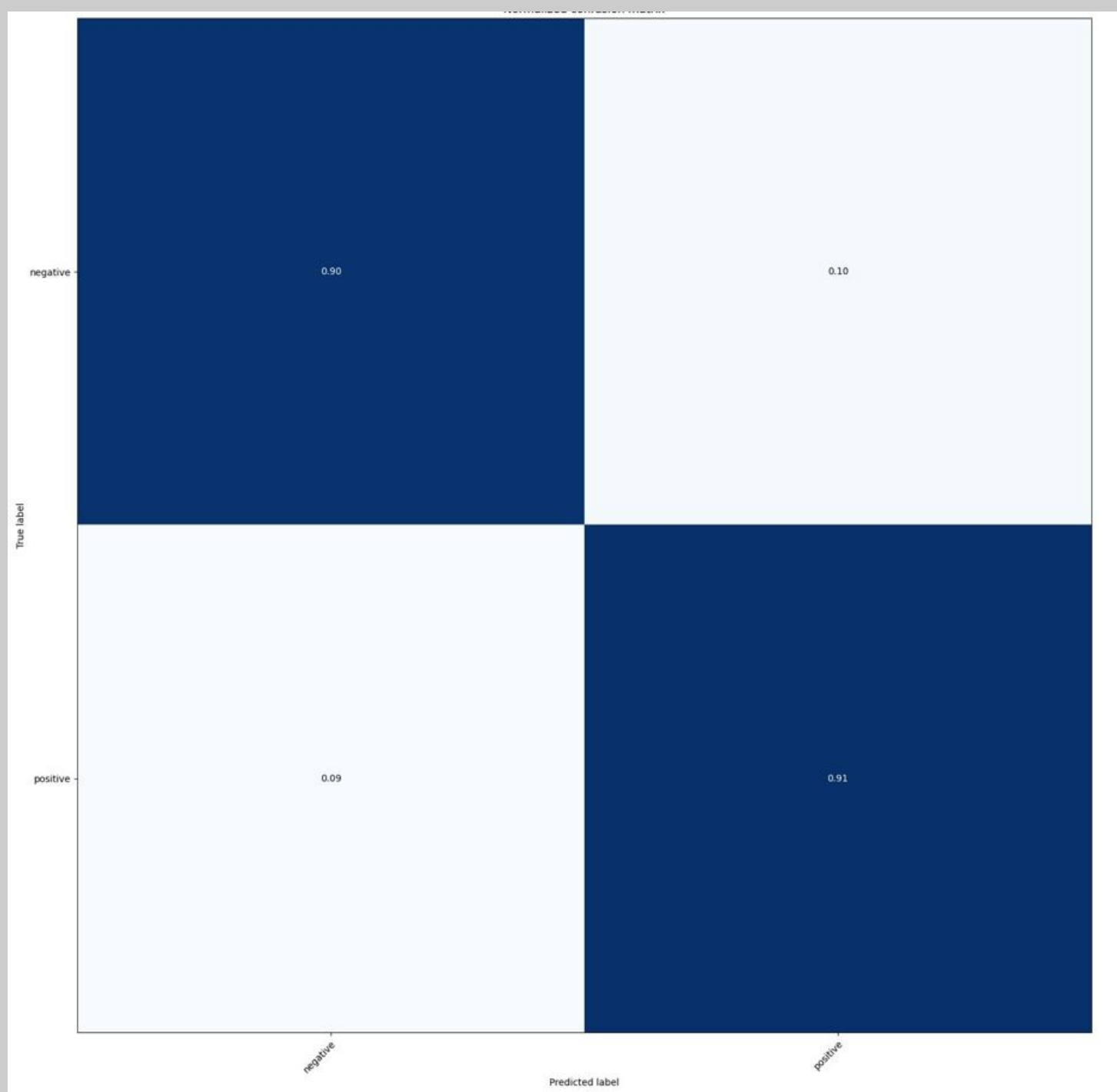
#Performance:

Seconds used for training: 183

Seconds used for classification: 19

#Classification report:

	precision	recall	f1-score	support
negative	0.91	0.90	0.90	4985
positive	0.90	0.91	0.90	5015
accuracy			0.90	10000
macro avg	0.90	0.90	0.90	10000
weighted avg	0.90	0.90	0.90	10000



Vor und Nachteile

Vorteile:

1. sehr gut für die Klassifikationen, die klare Grenzen haben.
2. Speicherplatz sparen
3. die Arbeit in hohen Dimensionen

Nachteile:

1. Zeitaufwendig bei unterschiedlichen Arten, zb Hund und Katze
2. Keine Hyperebene im niederdimensionalen Raum finden, möglich braucht es hohe Dimensionen

Entwicklung

Die Entwicklung von SVM :
Gesichtserkennung, Bioinformatik, Textklassifikation.

Textklassifikation :
Es kann die Worte in hohen Dimensionen bearbeiten und
die support Vektoren mit wichtigen Information finden.

