Unternehmensklassifikation

anhand von gecrawlten Snippets

Author

STEPHAN LENERT

PETER BEHRENS

TOM MÜLLER

Date

06.07.2021

Agenda

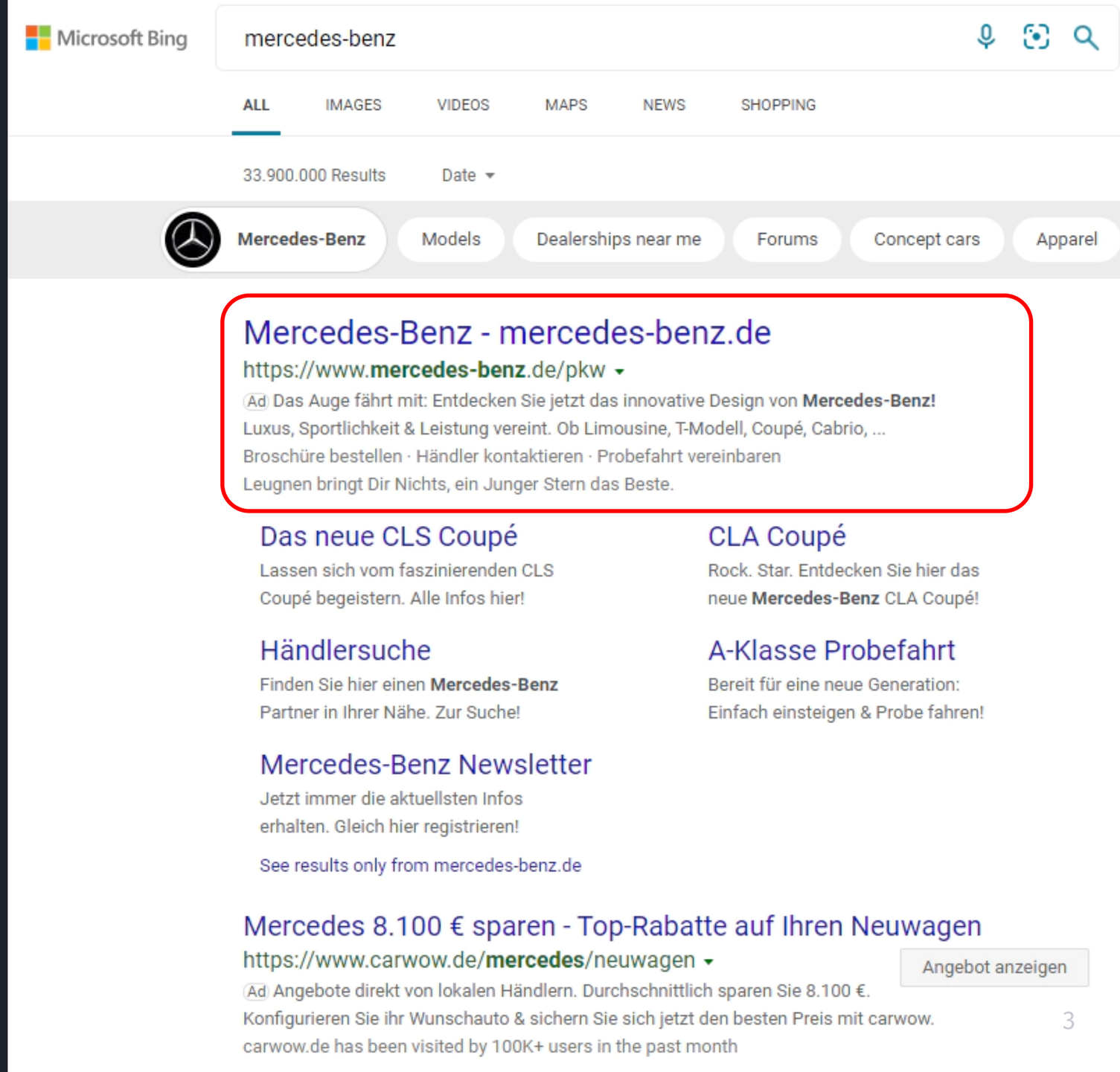
- Ziel des Projektes
- Datenquellen & Transformation
- Datenverteilung
- Theorie des Lösungsansatz
- Klassifikationsalgorithmen
- Herausforderungen
- Ergebnisse
- Verbesserungsansatz
- Demo

Was war das Ziel?

AUSGANGSSITUATION

Klassifikation der Unternehmen in verschiedene Segmente mithilfe von Snippets der Suchmaschine Bing.

Python Crawler zum Download der Snippets.



DATENQUELLEN & -TRANSFORMATION

Datenherkunft

Kaggle: Forbes Top 2000 Companies

Crawlen

Implementieren des Crawlers in Python und Crawlen des Datensatzes

Manuelle Datenbereinigung

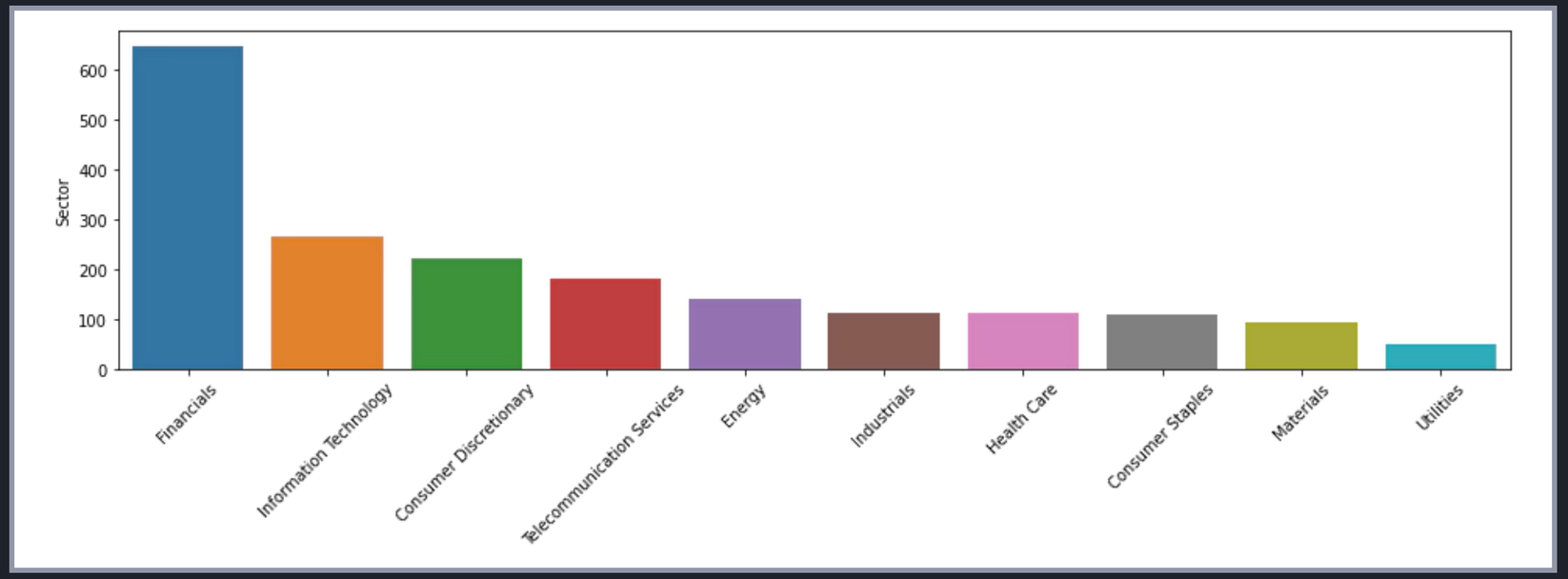
Englische und nichtssagende Kurzbeschreibungen wurden durch bessere ersetzt

Datenvorverarbeitung

Entfernen von Großbuchstaben, Satzzeichen & Stoppwörtern

Datenverteilung

"Financials" ist überrepräsentiert im Datensatz. Alle anderen Sektoren sind ausgeglichen.



Theorie des Lösungsansatz

1. DATEN-VORVERARBEITUNG

```
def prepare_snippets(snippet, raw_string_return = False, remove_int = True, lowercase = True, stopwords =
True, punctuations = True, only_nouns_n_adjs = True, lammatize = True, reduce = True, word_embeddings = True):
...
```

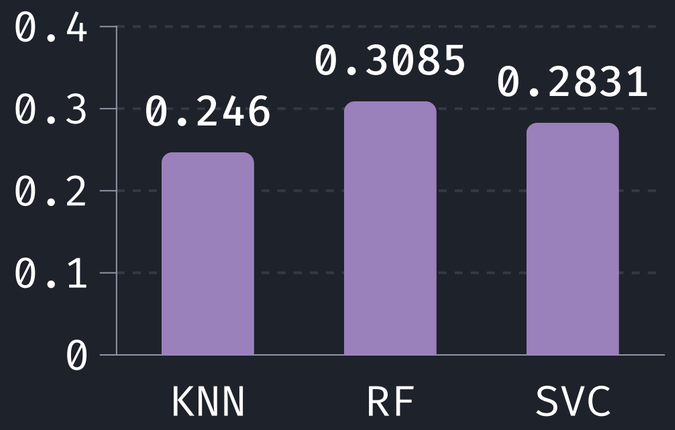
2. MODEL HYPERPARAMETER TUNING

```
from sklearn.model_selection import RandomizedSearchCV
...
search = RandomizedSearchCV(rf, space, n_iter=500, scoring='accuracy', n_jobs=-1, cv=cv, verbose = 1)
```

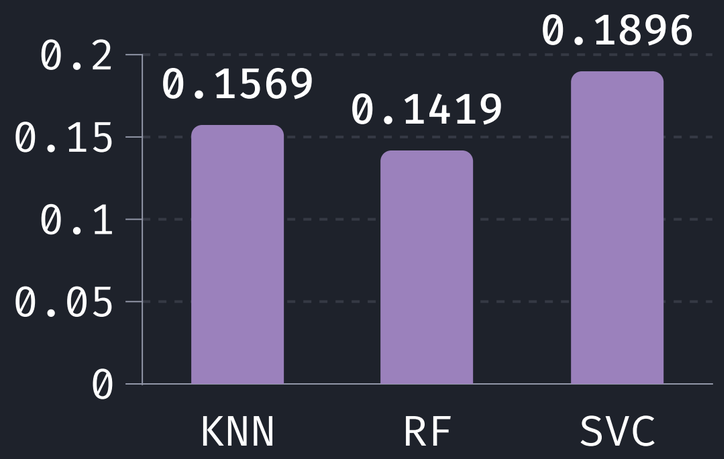
3. STACKING + CROSSVALIDATION

```
from sklearn.ensemble import StackingClassifier
stacked_clf = StackingClassifier(estimators=classifiers, final_estimator=LogisticRegression(), cv = 10 )
```

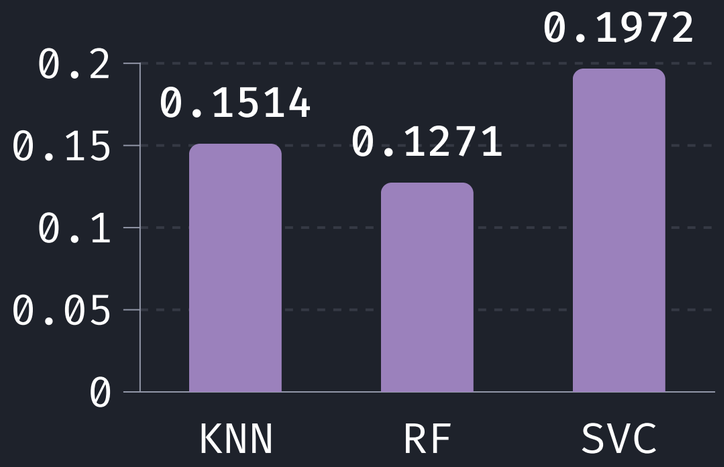
VERWENDETE KLASSIFIKATIONSALGORITHMEN



Precision



Recall



F1

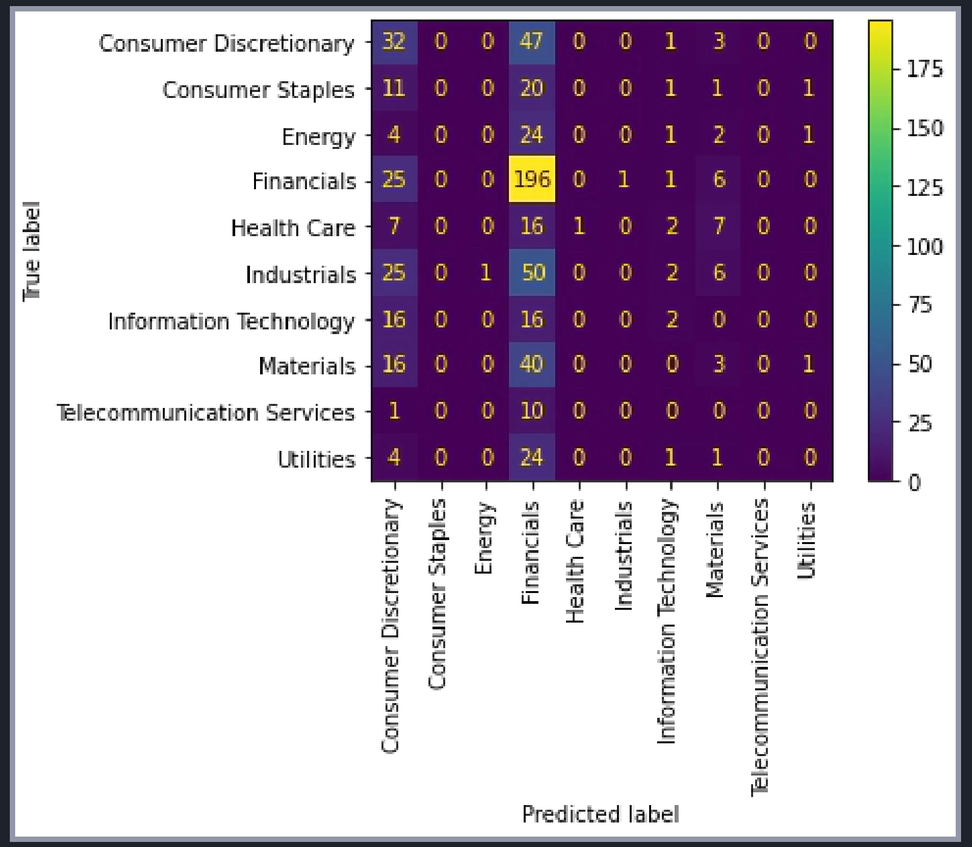
```
knn = Pipeline([('scaler', StandardScaler()),('clf',KNeighborsClassifier(weights="distance", n_neighbors=22))])
rf = Pipeline([('scaler', StandardScaler()),('clf',RandomForestClassifier(n_estimators = 200, criterion = "gini", max_depth = 10,
      max_features = "auto", min_samples_leaf = 0.005,
      min_samples_split = 0.005, n_jobs = -1, random_state = 1000, bootstrap = False))])
svc = Pipeline([('scaler', StandardScaler()),('clf',SVC(C = 12.450980939125662, degree = 4, gamma = 0.008983921439908003,
      kernel = "rbf", probability = True, class_weight=None))])
```

Unsere Herausforderungen:

- Datensatz
- Crawler
- Modell Performance Verbessern
- Deployment

Ergebnisse

- Vorhersage beruht auf finalem stacked Modell auf gesamten Testdatensatz
- Einsehen und Klassifizierung der Unternehmenssnippets in der App



Test Set Accuracy Score

0,3714

Train Set Accuracy Score

0,705

Macro Average Precison

0,20

Macro Average Recall

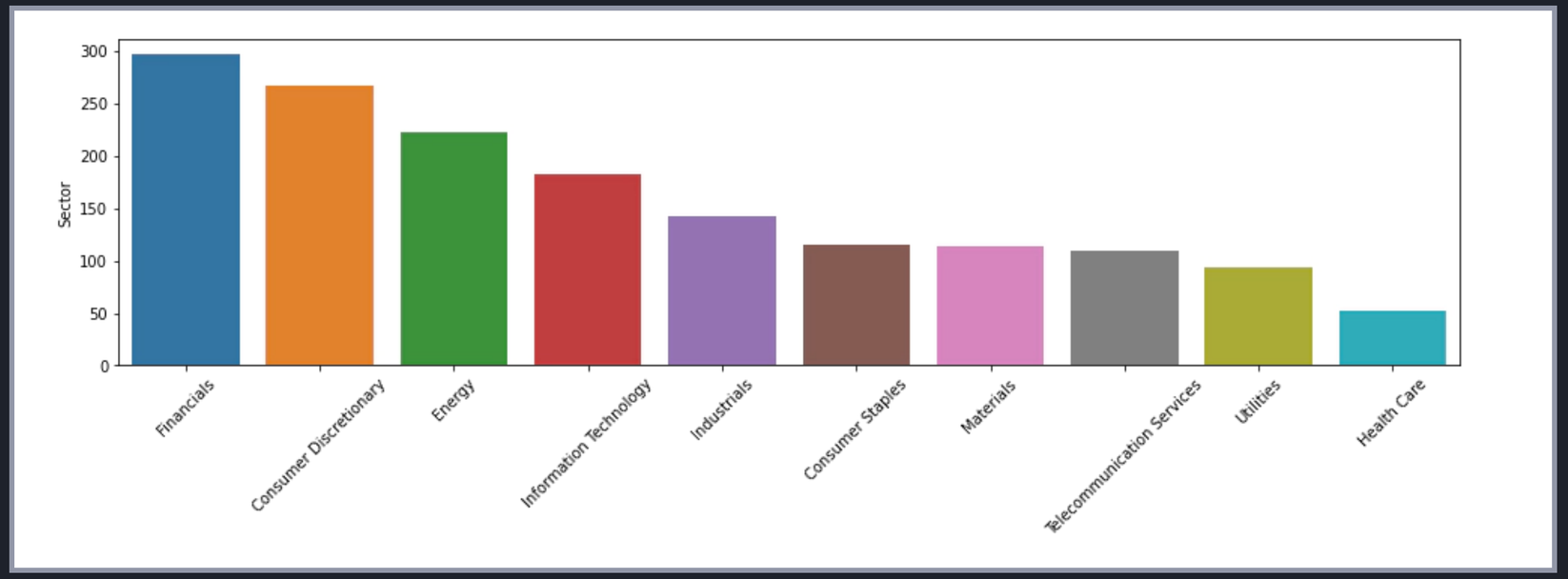
0,14

Macro Average F1-Score

0,11

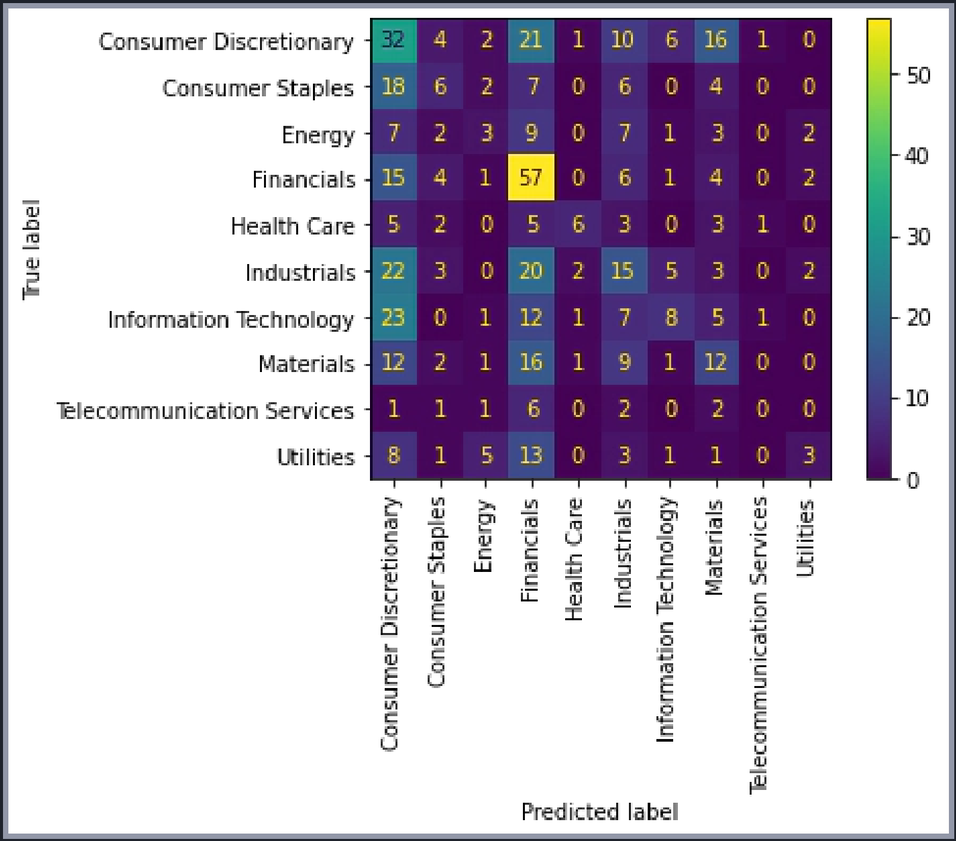
Datenverteilung - Verbesserung

"Financials" wurde angepasst, sodass der Datensatz ausgeglichener ist.



Ergebnisse

- Nach Ausgleich des Datensatzes
- Finales Model trainiert auf ganzem Datensatz



Test Set Accuracy Score

0,2747

Train Set Accuracy Score

0,8953

Macro Average Precison

0,27

Macro Average Recall

0,21

Macro Average F1-Score

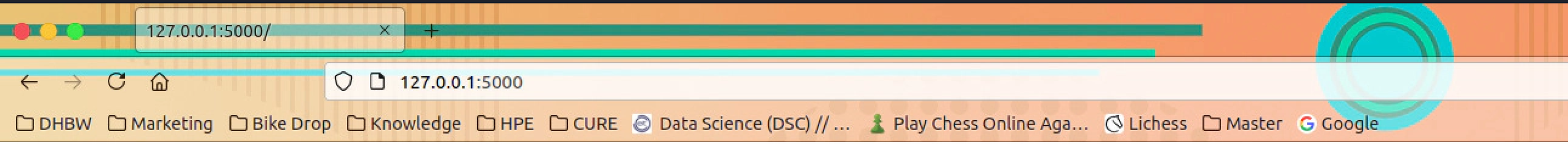
0,21

</>

Demo

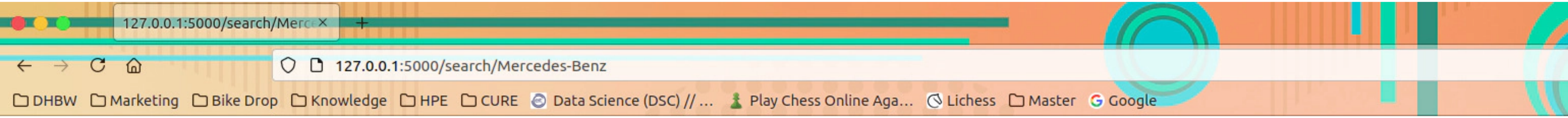
LOCAL FLASK APP

Mercedes-Benz



Enter Company

Mercedes-Benz | Daten absenden



Company:

Mercedes-Benz

Search-Snippet:

https://www.mercedes-benz.deLuxus, Sportlichkeit & Leistung vereint. Ob Limousine, T-Modell, Coupé, Cabrio, Roadster, SUV & mehr. Erleben Sie die Produkte von Mercedes-Benz.

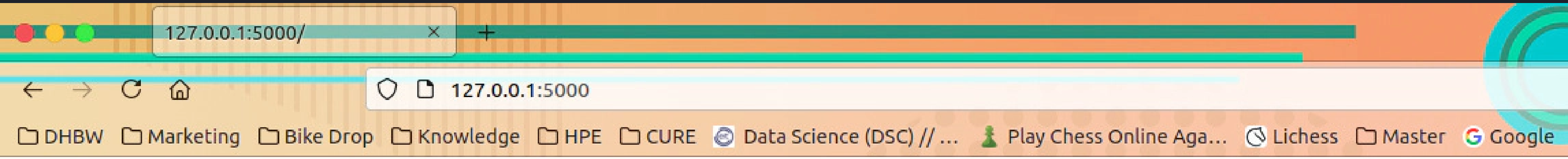
Predicted Sector:

Consumer Discretionary

Pitch

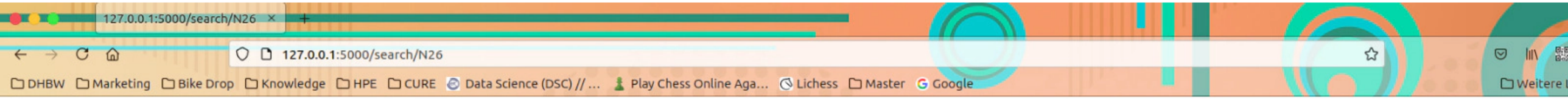
Zurück

N26



Enter Company

N26 Daten absenden



Company:

N26

Search-Snippet:

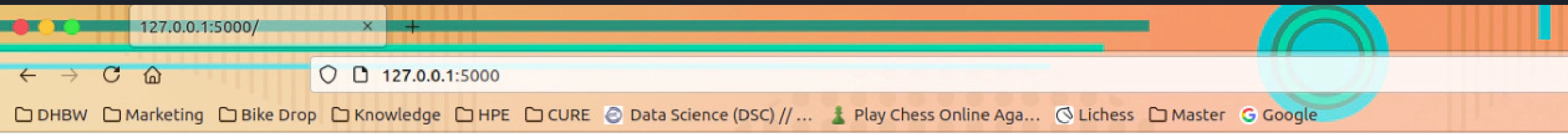
https://n26.com/de-deN26 ist die mobile Bank. Verwalte dein Girokonto 100 % mobil und verfolge deine Einnahmen und Ausgaben übersichtlich in Echtzeit. Eröffne jetzt dein N26 Girokonto in wenigen Minuten auf deinem Smartphone und nutze es, schon bevor deine physische Karte eintrifft. Girokonto eröffnen.

Predicted Sector:

Financials

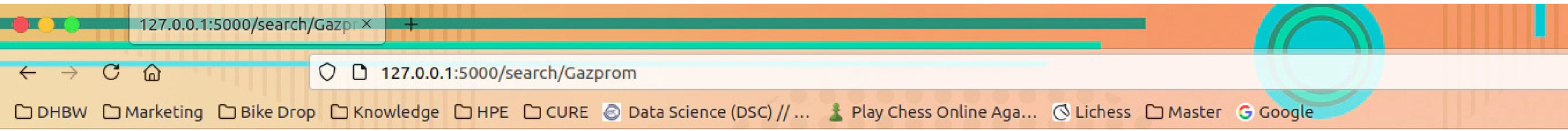
Zurück

Gazprom



Enter Company

Gazprom | Daten absenden



Company:

Gazprom

Search-Snippet:

GAZPROM Germania GmbH. Markgrafenstraße 23 . 10117 Berlin. Telefon +49 (0)30 20195-0. Telefax +49 (0)30 20195-313. E-Mail: info@gazprom-germania.de –

Predicted Sector:

Financials

Pitch

Zurück

ARBEITSTEILUNG

Name	Rolle	Aufgabe
Tom	Der Mann fürs Schöne	Datenverarbeitung, Präsentation und Dokumentation, Flask Frontend
Peter	Backend	Recherche (Algorithmen, Datenquellen), Datenvorverarbeitung, Crawler, Dokumentation
Stephan	Backend	Implementierung, Evaluierung und Verbesserung der Algorithmen, Aufsetzen der Flask-App

"teamwork makes the dream work"

</>

Vielen Dank für Ihre Aufmerksamkeit!

PETER, STEPHAN, TOM