Klasifikacija novinskih članaka

Tatjana Radovanović

Matematički fakultet, Univerzitet u Beogradu

September 9, 2019

Uvod

1. Podaci

2. Klasifikacija

Osnovno o podacima

- https://www.kaggle.com/rmisra/news-category-dataset
- 202372 reda
- Atributi
  - authors
  - category
  - date
  - headline
  - link
  - short_description

Pretprocesiranje podataka

- Atributi 'headline' i 'short_description' spojeni u atribut 'text'
- Uklonjeni specijalni znakovi iz teksta (.,_!? itd)
- Sva slova u atributu 'text' su konvertovana u mala
- Uklonjene su stop reči
- Uklonjeni su članci koji u tekstu imaju manje od 7 reči
- Za klasifikaciju su upotrebljeni samo atributi 'category' i 'text', ostali su uklonjeni

Pretprocesiranje podataka

- Podaci podeljeni u 41 kategoriju
- Srodne kategorije svrstane u jednu
  - 'THE WORLDPOST', 'WORLDPOST', **'WORLD NEWS'**
  - 'ARTS', 'CULTURE & ARTS', **'ARTS & CULTURE'**
  - 'COLLEGE', **'EDUCATION'**
  - 'PARENTING', **'PARENTS'**
  - 'STYLE', **'STYLE & BEAUTY'**
  - 'GREEN', **'ENVIRONMENT'**
  - 'TASTE', **'FOOD & DRINK'**
  - 'WELLNESS', **'HEALTHY LIVING'**
- Preostala 31 kategorija

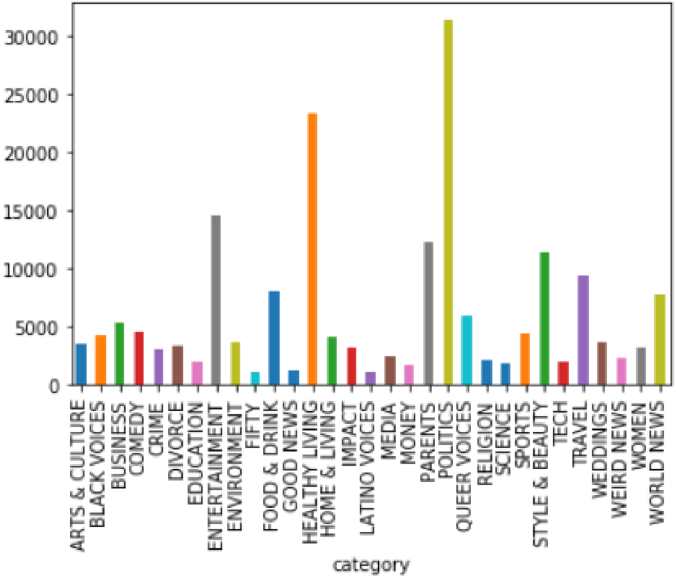


Figure: Broj članaka po kategorijama

Klasifikacija

- Ciljna promenljiva y je atribut 'category', a x je 'text'
- Podela na trening i test skup
- Stratifikacija po y
- Vektorizacija reči
    - Binarni oblik - 5000 najfrekventnijih reči
    - CountVectorizer - 5000 najfrekventnijih reči
    - TF-IDF - 10250 najfrekventnijih reči

Algoritmi

- Naivni Bajesov klasifikator
- Multinomijalna logistička regresija
- Metod potpornih vektora
- Slučajne šume

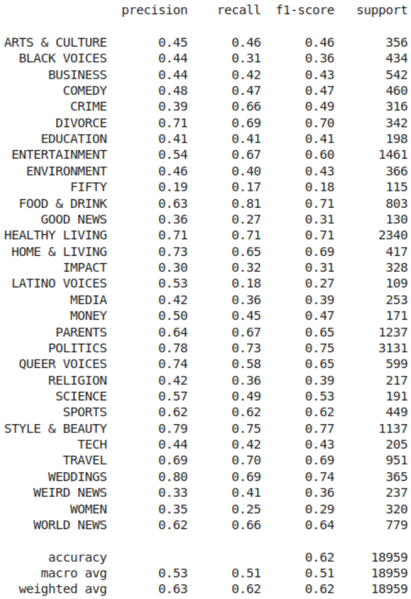
Naivni Bajesov klasifikator

- Unakrsna validacija
- Binarni oblik
  - Bernulijeva raspodela
  - Parametri najboljeg modela alpha=1, binarize=0.0, class_prior=None, fit_prior=True
- CountVectorizer
  - Multinomijalna raspodela
  - Parametri najboljeg modela alpha=1, class_prior=None, fit_prior=True
- TF-IDF
  - Multinomijalna raspodela
  - Parametri najboljeg modela alpha=1, class_prior=None, fit_prior=False

Naivni Bajesov klasifikator

Table: Preciznost za Naivni Bajesov klasifikator

	Trening skup	Validacioni skup	Test skup
Binarni oblik	0.6614	0.6136	0.6215
CountVectorizer	0.6637	0.6171	0.6215
TF-IDF	0.6614	0.6396	0.6397

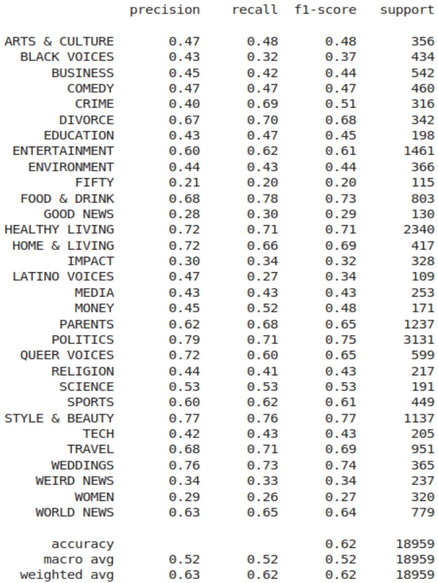


```
                precision    recall  f1-score   support

ARTS & CULTURE       0.45      0.46      0.46       356
  BLACK VOICES       0.44      0.31      0.36       434
      BUSINESS       0.44      0.42      0.43       542
        COMEDY       0.48      0.47      0.47       460
         CRIME       0.39      0.66      0.49       316
       DIVORCE       0.71      0.69      0.70       342
     EDUCATION       0.41      0.41      0.41       198
 ENTERTAINMENT       0.54      0.67      0.60      1461
   ENVIRONMENT       0.46      0.40      0.43       366
         FIFTY       0.19      0.17      0.18       115
  FOOD & DRINK       0.63      0.81      0.71       803
     GOOD NEWS       0.36      0.27      0.31       130
HEALTHY LIVING       0.71      0.71      0.71      2340
 HOME & LIVING       0.73      0.65      0.69       417
        IMPACT       0.30      0.32      0.31       328
 LATINO VOICES       0.53      0.18      0.27       109
         MEDIA       0.42      0.36      0.39       253
         MONEY       0.50      0.45      0.47       171
       PARENTS       0.64      0.67      0.65      1237
      POLITICS       0.78      0.73      0.75      3131
  QUEER VOICES       0.74      0.58      0.65       599
      RELIGION       0.42      0.36      0.39       217
       SCIENCE       0.57      0.49      0.53       191
        SPORTS       0.62      0.62      0.62       449
STYLE & BEAUTY       0.79      0.75      0.77      1137
          TECH       0.44      0.42      0.43       205
        TRAVEL       0.69      0.70      0.69       951
      WEDDINGS       0.80      0.69      0.74       365
    WEIRD NEWS       0.33      0.41      0.36       237
         WOMEN       0.35      0.25      0.29       320
    WORLD NEWS       0.62      0.66      0.64       779

      accuracy                           0.62     18959
     macro avg       0.53      0.51      0.51     18959
  weighted avg       0.63      0.62      0.62     18959
```

Figure: Izveštaj klasifikacije Naivnog Bajesovog klasifikatora za binarni oblik vektorizacije reči

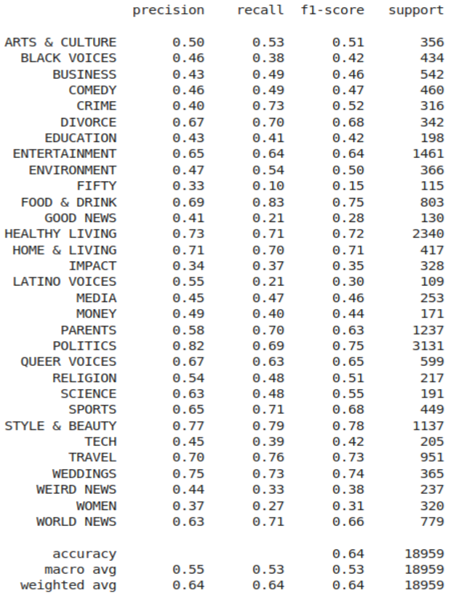


```
                  precision    recall  f1-score   support

 ARTS & CULTURE       0.47      0.48      0.48       356
   BLACK VOICES       0.43      0.32      0.37       434
       BUSINESS       0.45      0.42      0.44       542
         COMEDY       0.47      0.47      0.47       460
          CRIME       0.40      0.69      0.51       316
        DIVORCE       0.67      0.70      0.68       342
      EDUCATION       0.43      0.47      0.45       198
  ENTERTAINMENT       0.60      0.62      0.61      1461
    ENVIRONMENT       0.44      0.43      0.44       366
          FIFTY       0.21      0.20      0.20       115
   FOOD & DRINK       0.68      0.78      0.73       803
      GOOD NEWS       0.28      0.30      0.29       130
 HEALTHY LIVING       0.72      0.71      0.71      2340
  HOME & LIVING       0.72      0.66      0.69       417
         IMPACT       0.30      0.34      0.32       328
  LATINO VOICES       0.47      0.27      0.34       109
          MEDIA       0.43      0.43      0.43       253
          MONEY       0.45      0.52      0.48       171
        PARENTS       0.62      0.68      0.65      1237
       POLITICS       0.79      0.71      0.75      3131
   QUEER VOICES       0.72      0.60      0.65       599
       RELIGION       0.44      0.41      0.43       217
        SCIENCE       0.53      0.53      0.53       191
         SPORTS       0.60      0.62      0.61       449
 STYLE & BEAUTY       0.77      0.76      0.77      1137
           TECH       0.42      0.43      0.43       205
         TRAVEL       0.68      0.71      0.69       951
       WEDDINGS       0.76      0.73      0.74       365
     WEIRD NEWS       0.34      0.33      0.34       237
          WOMEN       0.29      0.26      0.27       320
     WORLD NEWS       0.63      0.65      0.64       779

       accuracy                           0.62     18959
      macro avg       0.52      0.52      0.52     18959
   weighted avg       0.63      0.62      0.62     18959
```

Figure: Izveštaj klasifikacije Naivnog Bajesovog klasifikatora za CountVectorizer



```
               precision    recall  f1-score   support

ARTS & CULTURE       0.50      0.53      0.51       356
  BLACK VOICES       0.46      0.38      0.42       434
      BUSINESS       0.43      0.49      0.46       542
        COMEDY       0.46      0.49      0.47       460
         CRIME       0.40      0.73      0.52       316
       DIVORCE       0.67      0.70      0.68       342
     EDUCATION       0.43      0.41      0.42       198
 ENTERTAINMENT       0.65      0.64      0.64      1461
   ENVIRONMENT       0.47      0.54      0.50       366
         FIFTY       0.33      0.10      0.15       115
  FOOD & DRINK       0.69      0.83      0.75       803
     GOOD NEWS       0.41      0.21      0.28       130
HEALTHY LIVING       0.73      0.71      0.72      2340
 HOME & LIVING       0.71      0.70      0.71       417
        IMPACT       0.34      0.37      0.35       328
 LATINO VOICES       0.55      0.21      0.30       109
         MEDIA       0.45      0.47      0.46       253
         MONEY       0.49      0.40      0.44       171
       PARENTS       0.58      0.70      0.63      1237
      POLITICS       0.82      0.69      0.75      3131
  QUEER VOICES       0.67      0.63      0.65       599
      RELIGION       0.54      0.48      0.51       217
       SCIENCE       0.63      0.48      0.55       191
        SPORTS       0.65      0.71      0.68       449
STYLE & BEAUTY       0.77      0.79      0.78      1137
          TECH       0.45      0.39      0.42       205
        TRAVEL       0.70      0.76      0.73       951
      WEDDINGS       0.75      0.73      0.74       365
    WEIRD NEWS       0.44      0.33      0.38       237
         WOMEN       0.37      0.27      0.31       320
    WORLD NEWS       0.63      0.71      0.66       779

      accuracy                           0.64     18959
     macro avg       0.55      0.53      0.53     18959
  weighted avg       0.64      0.64      0.64     18959
```

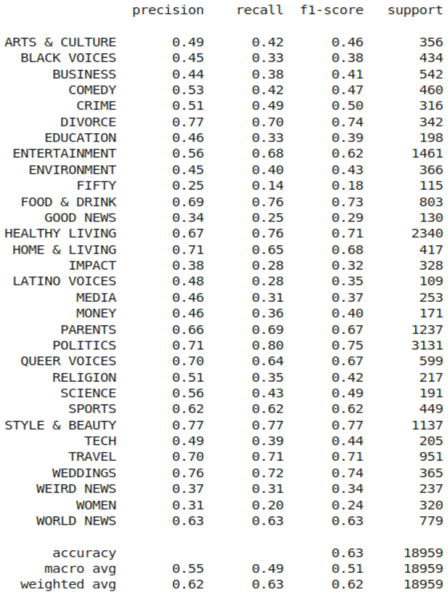
Figure: Izveštaj klasifikacije Naivnog Bajesovog klasifikatora za TF-IDF

Multinomijalna logistička regresija

- Unakrsna validacija
- CountVectorizer
- TF-IDF

Table: Preciznost za Multinomijalnu logističku regresiju

	Trening skup	Validacioni skup	Test skup
CountVectorizer	0.7590	0.6248	0.6328
TF-IDF	0.7329	0.6609	0.6672

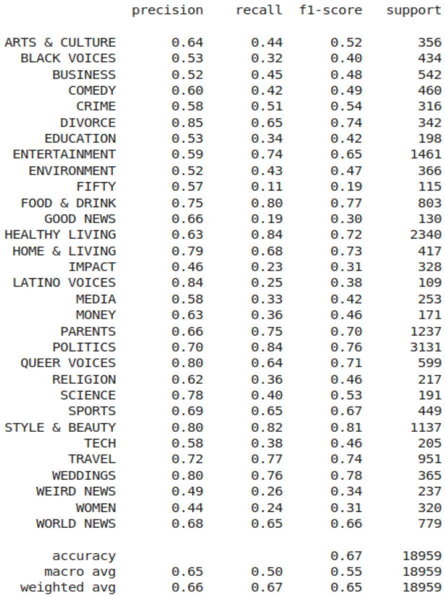


```
                precision    recall  f1-score   support

ARTS & CULTURE       0.49      0.42      0.46       356
  BLACK VOICES       0.45      0.33      0.38       434
      BUSINESS       0.44      0.38      0.41       542
        COMEDY       0.53      0.42      0.47       460
         CRIME       0.51      0.49      0.50       316
       DIVORCE       0.77      0.70      0.74       342
     EDUCATION       0.46      0.33      0.39       198
 ENTERTAINMENT       0.56      0.68      0.62      1461
   ENVIRONMENT       0.45      0.40      0.43       366
         FIFTY       0.25      0.14      0.18       115
  FOOD & DRINK       0.69      0.76      0.73       803
     GOOD NEWS       0.34      0.25      0.29       130
HEALTHY LIVING       0.67      0.76      0.71      2340
 HOME & LIVING       0.71      0.65      0.68       417
        IMPACT       0.38      0.28      0.32       328
 LATINO VOICES       0.48      0.28      0.35       109
         MEDIA       0.46      0.31      0.37       253
         MONEY       0.46      0.36      0.40       171
       PARENTS       0.66      0.69      0.67      1237
      POLITICS       0.71      0.80      0.75      3131
  QUEER VOICES       0.70      0.64      0.67       599
      RELIGION       0.51      0.35      0.42       217
       SCIENCE       0.56      0.43      0.49       191
        SPORTS       0.62      0.62      0.62       449
STYLE & BEAUTY       0.77      0.77      0.77      1137
          TECH       0.49      0.39      0.44       205
        TRAVEL       0.70      0.71      0.71       951
      WEDDINGS       0.76      0.72      0.74       365
    WEIRD NEWS       0.37      0.31      0.34       237
         WOMEN       0.31      0.20      0.24       320
    WORLD NEWS       0.63      0.63      0.63       779

      accuracy                           0.63     18959
     macro avg       0.55      0.49      0.51     18959
  weighted avg       0.62      0.63      0.62     18959
```

Figure: Izveštaj klasifikacije Multinomijalne logističke regresije za CountVectorizer



```
                precision    recall  f1-score   support

ARTS & CULTURE       0.64      0.44      0.52       356
  BLACK VOICES       0.53      0.32      0.40       434
      BUSINESS       0.52      0.45      0.48       542
        COMEDY       0.60      0.42      0.49       460
         CRIME       0.58      0.51      0.54       316
       DIVORCE       0.85      0.65      0.74       342
     EDUCATION       0.53      0.34      0.42       198
 ENTERTAINMENT       0.59      0.74      0.65      1461
   ENVIRONMENT       0.52      0.43      0.47       366
         FIFTY       0.57      0.11      0.19       115
  FOOD & DRINK       0.75      0.80      0.77       803
     GOOD NEWS       0.66      0.19      0.30       130
HEALTHY LIVING       0.63      0.84      0.72      2340
 HOME & LIVING       0.79      0.68      0.73       417
        IMPACT       0.46      0.23      0.31       328
 LATINO VOICES       0.84      0.25      0.38       109
         MEDIA       0.58      0.33      0.42       253
         MONEY       0.63      0.36      0.46       171
       PARENTS       0.66      0.75      0.70      1237
      POLITICS       0.70      0.84      0.76      3131
  QUEER VOICES       0.80      0.64      0.71       599
      RELIGION       0.62      0.36      0.46       217
       SCIENCE       0.78      0.40      0.53       191
        SPORTS       0.69      0.65      0.67       449
STYLE & BEAUTY       0.80      0.82      0.81      1137
          TECH       0.58      0.38      0.46       205
        TRAVEL       0.72      0.77      0.74       951
      WEDDINGS       0.80      0.76      0.78       365
    WEIRD NEWS       0.49      0.26      0.34       237
         WOMEN       0.44      0.24      0.31       320
    WORLD NEWS       0.68      0.65      0.66       779

      accuracy                           0.67     18959
     macro avg       0.65      0.50      0.55     18959
  weighted avg       0.66      0.67      0.65     18959
```

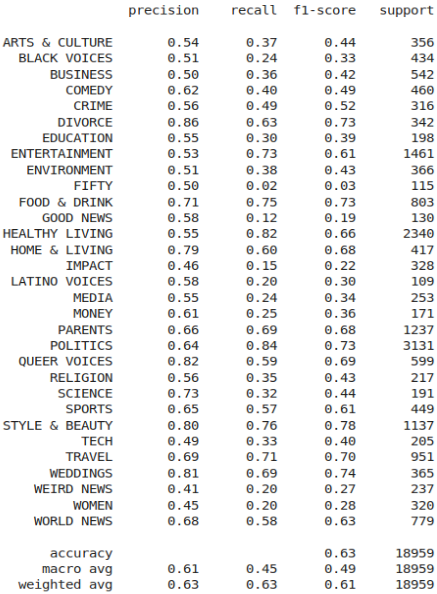
Figure: Izveštaj klasifikacije Multinomijalne logističke regresije za TF-IDF

Metod potpornih vektora

- CountVectorizer
- rbf kernel
- 'one-vs-rest'

Table: Preciznost za Metod potpornih vektora

	Trening skup	Test skup
CountVectorizer	0.8543	0.6265



```
                precision    recall  f1-score   support

ARTS & CULTURE       0.54      0.37      0.44       356
  BLACK VOICES       0.51      0.24      0.33       434
      BUSINESS       0.50      0.36      0.42       542
        COMEDY       0.62      0.40      0.49       460
         CRIME       0.56      0.49      0.52       316
       DIVORCE       0.86      0.63      0.73       342
     EDUCATION       0.55      0.30      0.39       198
 ENTERTAINMENT       0.53      0.73      0.61      1461
   ENVIRONMENT       0.51      0.38      0.43       366
         FIFTY       0.50      0.02      0.03       115
  FOOD & DRINK       0.71      0.75      0.73       803
     GOOD NEWS       0.58      0.12      0.19       130
HEALTHY LIVING       0.55      0.82      0.66      2340
 HOME & LIVING       0.79      0.60      0.68       417
        IMPACT       0.46      0.15      0.22       328
 LATINO VOICES       0.58      0.20      0.30       109
         MEDIA       0.55      0.24      0.34       253
         MONEY       0.61      0.25      0.36       171
       PARENTS       0.66      0.69      0.68      1237
      POLITICS       0.64      0.84      0.73      3131
  QUEER VOICES       0.82      0.59      0.69       599
      RELIGION       0.56      0.35      0.43       217
       SCIENCE       0.73      0.32      0.44       191
        SPORTS       0.65      0.57      0.61       449
STYLE & BEAUTY       0.80      0.76      0.78      1137
          TECH       0.49      0.33      0.40       205
        TRAVEL       0.69      0.71      0.70       951
      WEDDINGS       0.81      0.69      0.74       365
    WEIRD NEWS       0.41      0.20      0.27       237
         WOMEN       0.45      0.20      0.28       320
    WORLD NEWS       0.68      0.58      0.63       779

      accuracy                           0.63     18959
     macro avg       0.61      0.45      0.49     18959
  weighted avg       0.63      0.63      0.61     18959
```

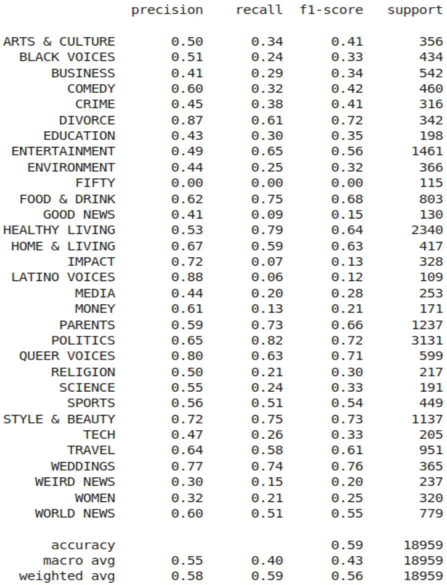
Figure: Izveštaj klasifikacije za metod potpornih vektora

Slučajne šume

- 100 estimatora
- Ginijev kriterijum podele

Table: Preciznost za slučajne šume

	Trening skup	Test skup
CountVectorizer	0.99993	0.58954
TF-IDF	0.99993	0.59550

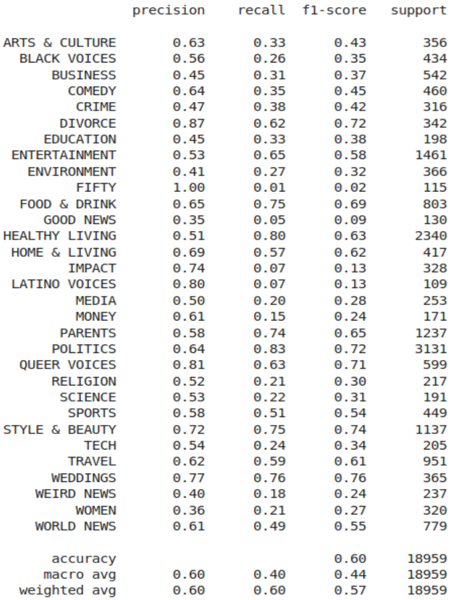


```
                precision    recall  f1-score   support

 ARTS & CULTURE       0.50      0.34      0.41       356
   BLACK VOICES       0.51      0.24      0.33       434
       BUSINESS       0.41      0.29      0.34       542
         COMEDY       0.60      0.32      0.42       460
          CRIME       0.45      0.38      0.41       316
        DIVORCE       0.87      0.61      0.72       342
      EDUCATION       0.43      0.30      0.35       198
  ENTERTAINMENT       0.49      0.65      0.56      1461
    ENVIRONMENT       0.44      0.25      0.32       366
          FIFTY       0.00      0.00      0.00       115
   FOOD & DRINK       0.62      0.75      0.68       803
      GOOD NEWS       0.41      0.09      0.15       130
 HEALTHY LIVING       0.53      0.79      0.64      2340
  HOME & LIVING       0.67      0.59      0.63       417
         IMPACT       0.72      0.07      0.13       328
  LATINO VOICES       0.88      0.06      0.12       109
          MEDIA       0.44      0.20      0.28       253
          MONEY       0.61      0.13      0.21       171
        PARENTS       0.59      0.73      0.66      1237
       POLITICS       0.65      0.82      0.72      3131
   QUEER VOICES       0.80      0.63      0.71       599
       RELIGION       0.50      0.21      0.30       217
        SCIENCE       0.55      0.24      0.33       191
         SPORTS       0.56      0.51      0.54       449
 STYLE & BEAUTY       0.72      0.75      0.73      1137
           TECH       0.47      0.26      0.33       205
         TRAVEL       0.64      0.58      0.61       951
       WEDDINGS       0.77      0.74      0.76       365
     WEIRD NEWS       0.30      0.15      0.20       237
          WOMEN       0.32      0.21      0.25       320
     WORLD NEWS       0.60      0.51      0.55       779

       accuracy                           0.59     18959
      macro avg       0.55      0.40      0.43     18959
   weighted avg       0.58      0.59      0.56     18959
```

Figure: Izveštaj klasifikacije slučajnih šuma za CountVectorizer



```
                precision    recall  f1-score   support

 ARTS & CULTURE       0.63      0.33      0.43       356
   BLACK VOICES       0.56      0.26      0.35       434
       BUSINESS       0.45      0.31      0.37       542
         COMEDY       0.64      0.35      0.45       460
          CRIME       0.47      0.38      0.42       316
        DIVORCE       0.87      0.62      0.72       342
      EDUCATION       0.45      0.33      0.38       198
  ENTERTAINMENT       0.53      0.65      0.58      1461
    ENVIRONMENT       0.41      0.27      0.32       366
          FIFTY       1.00      0.01      0.02       115
   FOOD & DRINK       0.65      0.75      0.69       803
      GOOD NEWS       0.35      0.05      0.09       130
 HEALTHY LIVING       0.51      0.80      0.63      2340
  HOME & LIVING       0.69      0.57      0.62       417
         IMPACT       0.74      0.07      0.13       328
  LATINO VOICES       0.80      0.07      0.13       109
          MEDIA       0.50      0.20      0.28       253
          MONEY       0.61      0.15      0.24       171
        PARENTS       0.58      0.74      0.65      1237
       POLITICS       0.64      0.83      0.72      3131
   QUEER VOICES       0.81      0.63      0.71       599
       RELIGION       0.52      0.21      0.30       217
        SCIENCE       0.53      0.22      0.31       191
         SPORTS       0.58      0.51      0.54       449
 STYLE & BEAUTY       0.72      0.75      0.74      1137
           TECH       0.54      0.24      0.34       205
         TRAVEL       0.62      0.59      0.61       951
       WEDDINGS       0.77      0.76      0.76       365
     WEIRD NEWS       0.40      0.18      0.24       237
          WOMEN       0.36      0.21      0.27       320
     WORLD NEWS       0.61      0.49      0.55       779

       accuracy                           0.60     18959
      macro avg       0.60      0.40      0.44     18959
   weighted avg       0.60      0.60      0.57     18959
```

Figure: Izveštaj klasifikacije slučajnih šuma za TF-IDF

Hvala na pažnji!