Università degli Studi di Milano - Bicocca

The 20 Newsgroups

di Beatrice Fumagalli e Matteo Porcino

Progetto di Text Mining and Search
Appello di gennaio

Dataset "20—newsgroups"

19997 documenti suddivisi in 20 newsgroups:

- alt.atheism
- comp.graphics
- comp.os.ms-windows.misc
- comp.sys.ibm.pc.hardware
- comp.sys.mac.hardware
- comp.windows.x
- misc.forsale
- rec.autos
- rec.motorcycles
- rec.sport.baseball
- rec.sport.hockey
- sci.crypt
- sci.electronics
- sci.med
- sci.space
- soc.religion.christian
- talk.politics.guns
- talk.politics.mideast
- talk.politics.misc
- talk.religion.misc

Text pre-processing

Dataset suddiviso in *Training set* (70%, 13997 articoli) *e Test set* (30%, 6000 articoli)

- Eliminazione dei metadata
- Eliminazione delle contrazioni
- Rimozione di URLs e indirizzi e-mail
- Tokenization
- Rimozione di righe vuote e tabs
- Rimozione della punteggiatura ed eventuali ulteriori spazi vuoti
- Eliminazione delle stringhe numeriche
- Normalizzazione dei caratteri in minuscolo
- Rimozione delle stop-words
- Eliminazione dei caratteri non unicode
- Eliminazione della punteggiatura
- Stemming e Lemmatization
- Eliminazione dei tokens di lunghezza inferiore a 2

Utilizzo della libreria NLTK

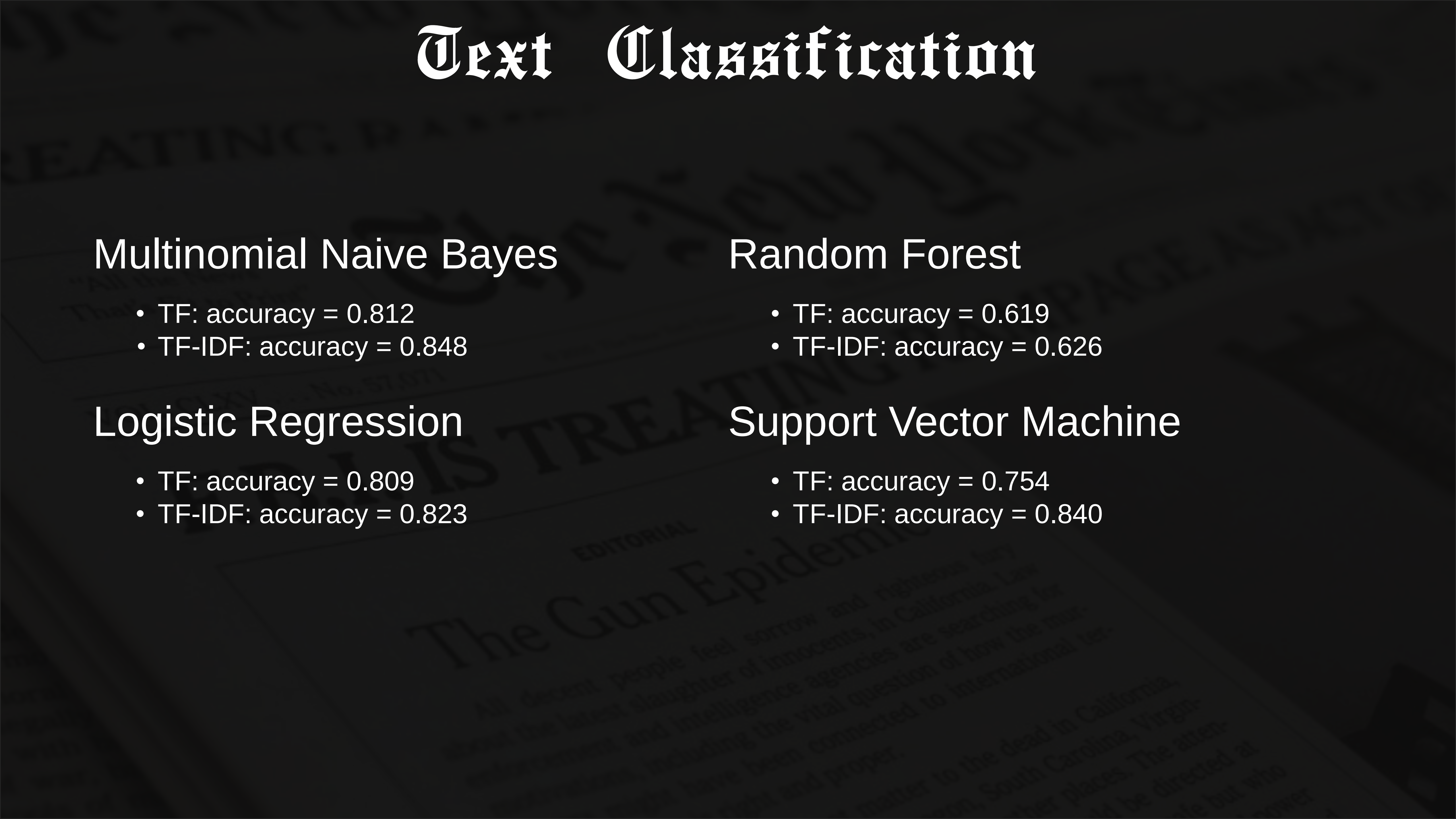
Text Representation

- **Term Frequencies (TF):** la Term Frequency $tf_{t,d}$ del termine t nel documento d è definito come il numero di volte che t si verifica in d.

- **Term Frequency times Inverse Document Frequency (TF-IDF):** il peso tf-idf di un termine è il prodotto del suo peso tf e del suo peso idf.

Utilizzo della libreria scikit-learn
Ciascuna rappresentazione fornisce una matrice Document-Term di dimensioni 13997x119613

Text Classification



Multinomial Naive Bayes

- TF: accuracy = 0.812
- TF-IDF: accuracy = 0.848

Logistic Regression

- TF: accuracy = 0.809
- TF-IDF: accuracy = 0.823

Random Forest

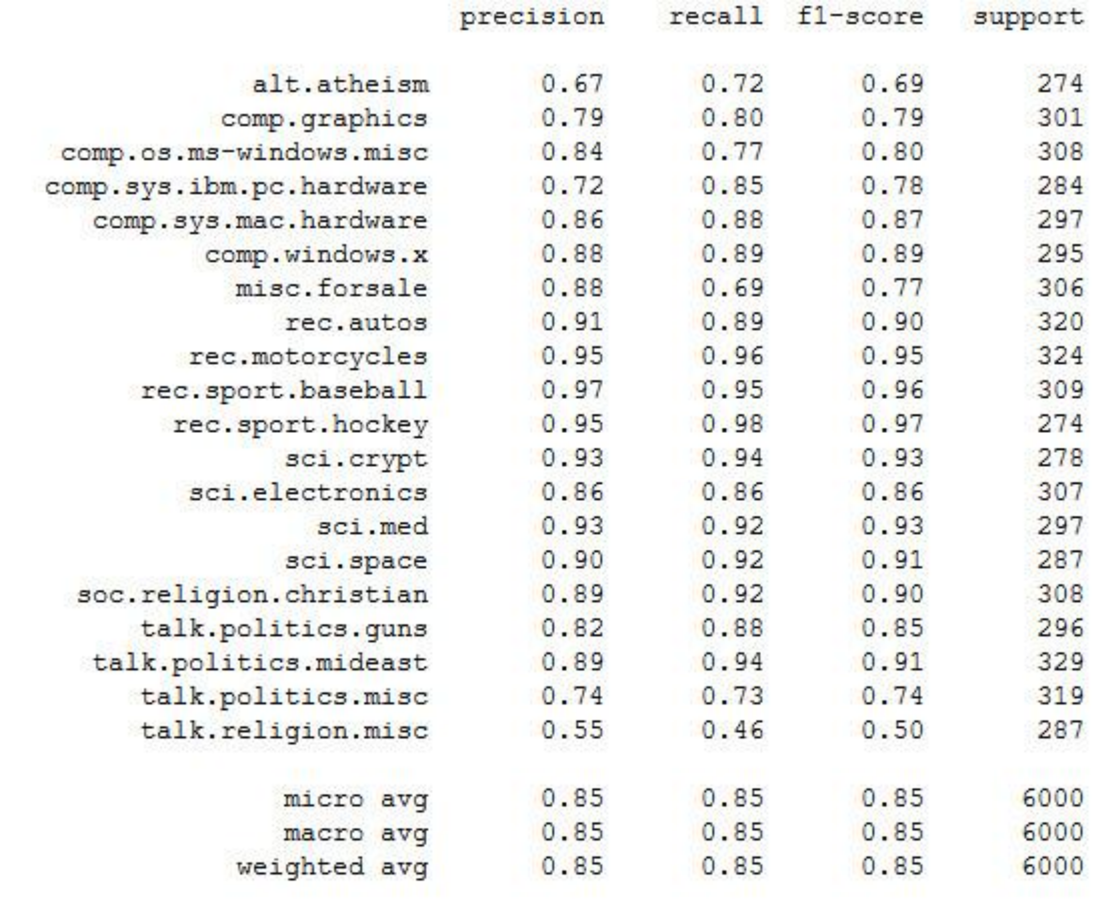
- TF: accuracy = 0.619
- TF-IDF: accuracy = 0.626

Support Vector Machine

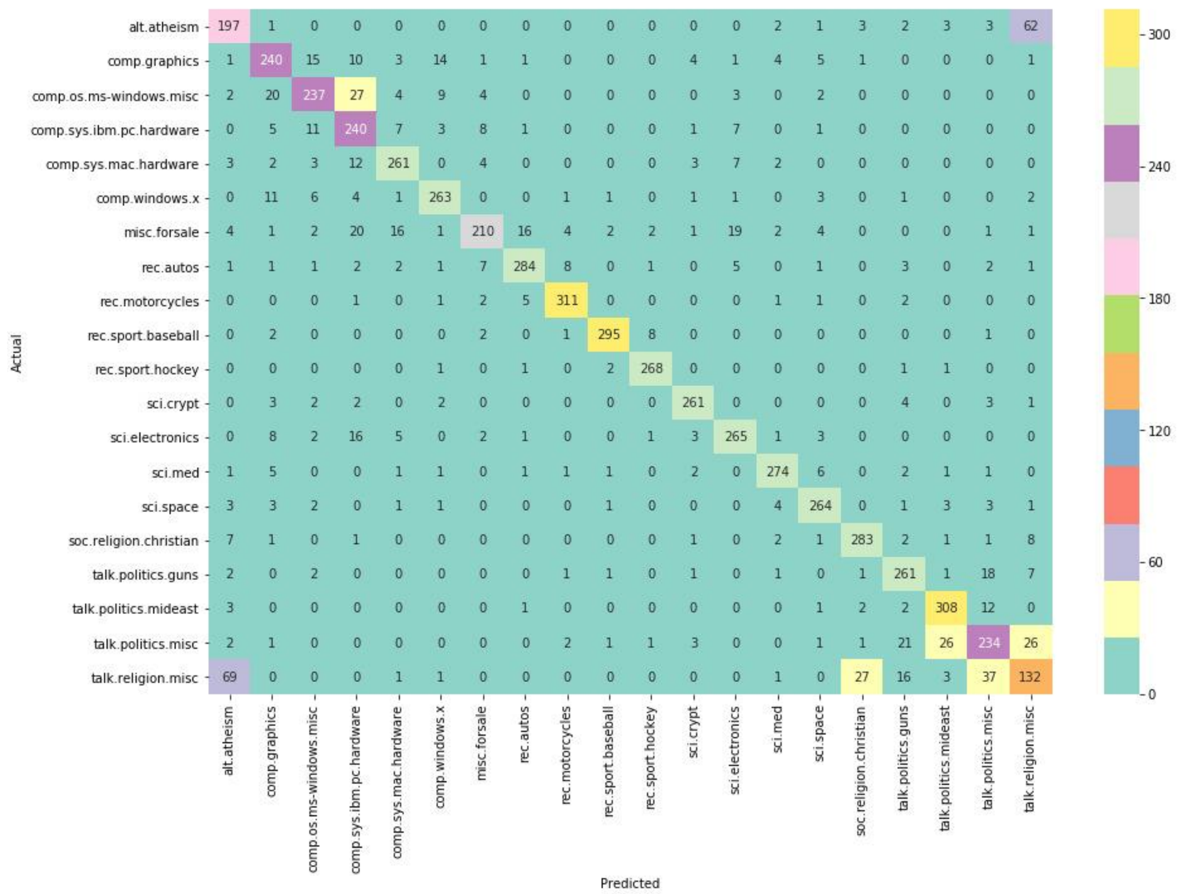
- TF: accuracy = 0.754
- TF-IDF: accuracy = 0.840

Risultati

Naive Bayes, TF-IDF:



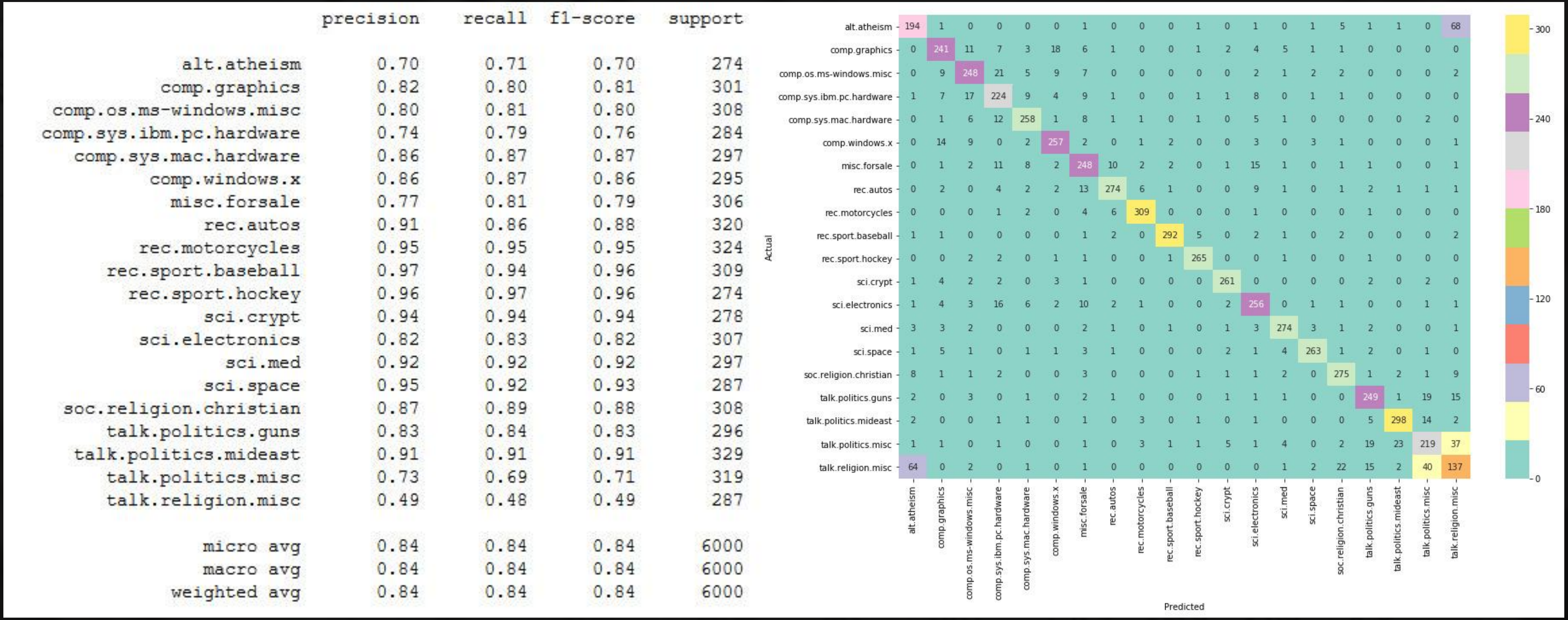
```
                          precision    recall  f1-score   support

             alt.atheism       0.67      0.72      0.69       274
           comp.graphics       0.79      0.80      0.79       301
 comp.os.ms-windows.misc       0.84      0.77      0.80       308
comp.sys.ibm.pc.hardware       0.72      0.85      0.78       284
   comp.sys.mac.hardware       0.86      0.88      0.87       297
          comp.windows.x       0.88      0.89      0.89       295
            misc.forsale       0.88      0.69      0.77       306
               rec.autos       0.91      0.89      0.90       320
         rec.motorcycles       0.95      0.96      0.95       324
      rec.sport.baseball       0.97      0.95      0.96       309
        rec.sport.hockey       0.95      0.98      0.97       274
               sci.crypt       0.93      0.94      0.93       278
         sci.electronics       0.86      0.86      0.86       307
                 sci.med       0.93      0.92      0.93       297
               sci.space       0.90      0.92      0.91       287
  soc.religion.christian       0.89      0.92      0.90       308
      talk.politics.guns       0.82      0.88      0.85       296
   talk.politics.mideast       0.89      0.94      0.91       329
      talk.politics.misc       0.74      0.73      0.74       319
      talk.religion.misc       0.55      0.46      0.50       287

               micro avg       0.85      0.85      0.85      6000
               macro avg       0.85      0.85      0.85      6000
            weighted avg       0.85      0.85      0.85      6000
```



	alt.atheism	comp.graphics	comp.os.ms-windows.misc	comp.sys.ibm.pc.hardware	comp.sys.mac.hardware	comp.windows.x	misc.forsale	rec.autos	rec.motorcycles	rec.sport.baseball	rec.sport.hockey	sci.crypt	sci.electronics	sci.med	sci.space	soc.religion.christian	talk.politics.guns	talk.politics.mideast	talk.politics.misc	talk.religion.misc
alt.atheism	197	1	0	0	0	0	0	0	0	0	0	0	0	2	1	3	2	3	3	62
comp.graphics	1	240	15	10	3	14	1	1	0	0	0	4	1	4	5	1	0	0	0	1
comp.os.ms-windows.misc	2	20	237	27	4	9	4	0	0	0	0	0	3	0	2	0	0	0	0	0
comp.sys.ibm.pc.hardware	0	5	11	240	7	3	8	1	0	0	0	1	7	0	1	0	0	0	0	0
comp.sys.mac.hardware	3	2	3	12	261	0	4	0	0	0	0	3	7	2	0	0	0	0	0	0
comp.windows.x	0	11	6	4	1	263	0	0	1	1	0	1	1	0	3	0	1	0	0	2
misc.forsale	4	1	2	20	16	1	210	16	4	2	2	1	19	2	4	0	0	0	1	1
rec.autos	1	1	1	2	2	1	7	284	8	0	1	0	5	0	1	0	3	0	2	1
rec.motorcycles	0	0	0	1	0	1	2	5	311	0	0	0	0	1	1	0	2	0	0	0
rec.sport.baseball	0	2	0	0	0	0	2	0	1	295	8	0	0	0	0	0	0	0	1	0
rec.sport.hockey	0	0	0	0	0	1	0	1	0	2	268	0	0	0	0	0	1	1	0	0
sci.crypt	0	3	2	2	0	2	0	0	0	0	0	261	0	0	0	0	4	0	3	1
sci.electronics	0	8	2	16	5	0	2	1	0	0	1	3	265	1	3	0	0	0	0	0
sci.med	1	5	0	0	1	1	0	1	1	1	0	2	0	274	6	0	2	1	1	0
sci.space	3	3	2	0	1	1	0	0	0	1	0	0	0	4	264	0	1	3	3	1
soc.religion.christian	7	1	0	1	0	0	0	0	0	0	0	1	0	2	1	283	2	1	1	8
talk.politics.guns	2	0	2	0	0	0	0	0	1	1	0	1	0	1	0	1	261	1	18	7
talk.politics.mideast	3	0	0	0	0	0	0	1	0	0	0	0	0	0	1	2	2	308	12	0
talk.politics.misc	2	1	0	0	0	0	0	0	2	1	1	3	0	0	1	1	21	26	234	26
talk.religion.misc	69	0	0	0	1	1	0	0	0	0	0	0	0	1	0	27	16	3	37	132

Risultati

Support Vector Machine, TF-IDF:



```
                          precision    recall  f1-score   support

             alt.atheism       0.70      0.71      0.70       274
           comp.graphics       0.82      0.80      0.81       301
 comp.os.ms-windows.misc       0.80      0.81      0.80       308
comp.sys.ibm.pc.hardware       0.74      0.79      0.76       284
   comp.sys.mac.hardware       0.86      0.87      0.87       297
          comp.windows.x       0.86      0.87      0.86       295
            misc.forsale       0.77      0.81      0.79       306
               rec.autos       0.91      0.86      0.88       320
         rec.motorcycles       0.95      0.95      0.95       324
      rec.sport.baseball       0.97      0.94      0.96       309
        rec.sport.hockey       0.96      0.97      0.96       274
               sci.crypt       0.94      0.94      0.94       278
         sci.electronics       0.82      0.83      0.82       307
                 sci.med       0.92      0.92      0.92       297
               sci.space       0.95      0.92      0.93       287
  soc.religion.christian       0.87      0.89      0.88       308
      talk.politics.guns       0.83      0.84      0.83       296
   talk.politics.mideast       0.91      0.91      0.91       329
      talk.politics.misc       0.73      0.69      0.71       319
      talk.religion.misc       0.49      0.48      0.49       287

               micro avg       0.84      0.84      0.84      6000
               macro avg       0.84      0.84      0.84      6000
            weighted avg       0.84      0.84      0.84      6000
```

	alt.atheism	comp.graphics	comp.os.ms-windows.misc	comp.sys.ibm.pc.hardware	comp.sys.mac.hardware	comp.windows.x	misc.forsale	rec.autos	rec.motorcycles	rec.sport.baseball	rec.sport.hockey	sci.crypt	sci.electronics	sci.med	sci.space	soc.religion.christian	talk.politics.guns	talk.politics.mideast	talk.politics.misc	talk.religion.misc
alt.atheism	194	1	0	0	0	0	1	0	0	0	1	0	1	0	1	5	1	1	0	68
comp.graphics	0	241	11	7	3	18	6	1	0	0	1	2	4	5	1	1	0	0	0	0
comp.os.ms-windows.misc	0	9	248	21	5	9	7	0	0	0	0	0	2	1	2	2	0	0	0	2
comp.sys.ibm.pc.hardware	1	7	17	224	9	4	9	1	0	0	1	1	8	0	1	1	0	0	0	0
comp.sys.mac.hardware	0	1	6	12	258	1	8	1	1	0	1	0	5	1	0	0	0	0	2	0
comp.windows.x	0	14	9	0	2	257	2	0	1	2	0	0	3	0	3	1	0	0	0	1
misc.forsale	0	1	2	11	8	2	248	10	2	2	0	1	15	1	0	1	1	0	0	1
rec.autos	0	2	0	4	2	2	13	274	6	1	0	0	9	1	0	1	2	1	1	1
rec.motorcycles	0	0	0	1	2	0	4	6	309	0	0	0	1	0	0	0	1	0	0	0
rec.sport.baseball	1	1	0	0	0	0	1	2	0	292	5	0	2	1	0	2	0	0	0	2
rec.sport.hockey	0	0	2	2	0	1	1	0	0	1	265	0	0	1	0	0	1	0	0	0
sci.crypt	1	4	2	2	0	3	1	0	0	0	0	261	0	0	0	0	2	0	2	0
sci.electronics	1	4	3	16	6	2	10	2	1	0	0	2	256	0	1	1	0	0	1	1
sci.med	3	3	2	0	0	0	2	1	0	1	0	1	3	274	3	1	2	0	0	1
sci.space	1	5	1	0	1	1	3	1	0	0	0	2	1	4	263	1	2	0	1	0
soc.religion.christian	8	1	1	2	0	0	3	0	0	0	1	1	1	2	0	275	1	2	1	9
talk.politics.guns	2	0	3	0	1	0	2	1	0	0	0	1	1	1	0	0	249	1	19	15
talk.politics.mideast	2	0	0	1	1	0	1	0	3	0	1	0	1	0	0	0	5	298	14	2
talk.politics.misc	1	1	0	1	0	0	1	0	3	1	1	5	1	4	0	2	19	23	219	37
talk.religion.misc	64	0	2	0	1	0	1	0	0	0	0	0	0	1	2	22	15	2	40	137