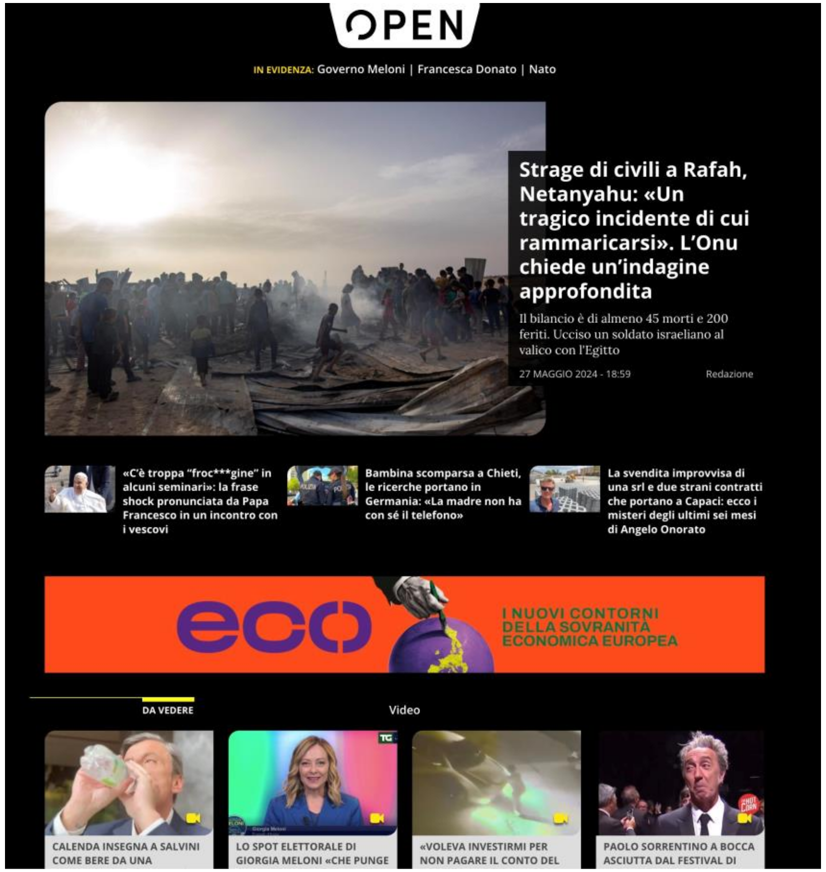
ARTICLES CATEGORIZER

DATA MINING AND MACHINE LEARNING
MIRKO DI LUCIA

UNIVERSITÀ DI PISA

INTRODUCTION

Find the most reliable data mining algorithm, and structure to categorize a subgroup of articles from a famous online newpaper website based on category



Strage di civili a Rafah, Netanyahu: «Un tragico incidente di cui rammaricarsi». L'Onu chiede un'indagine approfondita

Il bilancio è di almeno 45 morti e 200 feriti. Ucciso un soldato israeliano al valico con l'Egitto

27 MAGGIO 2024 - 18:59 Redazione

«C'è troppa "froc***gine" in alcuni seminari»: la frase shock pronunciata da Papa Francesco in un incontro con i vescovi

Bambina scomparsa a Chieti, le ricerche portano in Germania: «La madre non ha con sé il telefono»

La svendita improvvisa di una srl e due strani contratti che portano a Capaci: ecco i misteri degli ultimi sei mesi di Angelo Onorato

CALENDA INSEGNA A SALVINI COME BERE DA UNA

LO SPOT ELETTORALE DI GIORGIA MELONI «CHE PUNGE

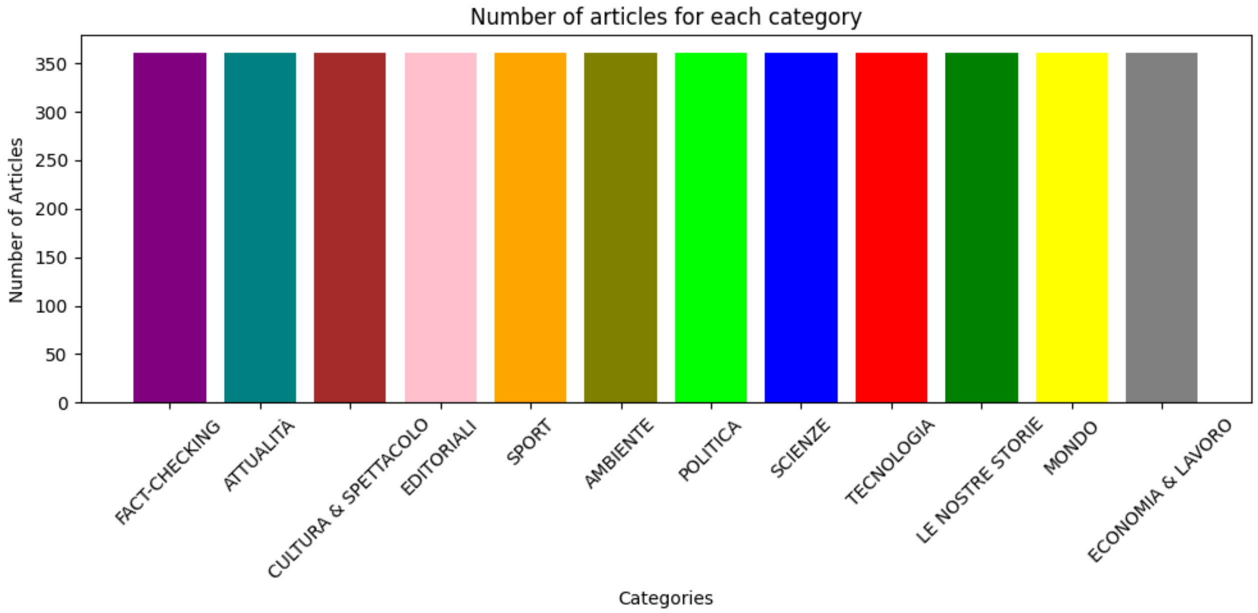
«VOLEVA INVESTIRMI PER NON PAGARE IL CONTO DEL

PAOLO SORRENTINO A BOCCA ASCIUTTA DAL FESTIVAL DI

DATASET DESCRIPTION

Dataset was scraped from the website

- Dataset is composed from 4332 articles
- Each article had a single category

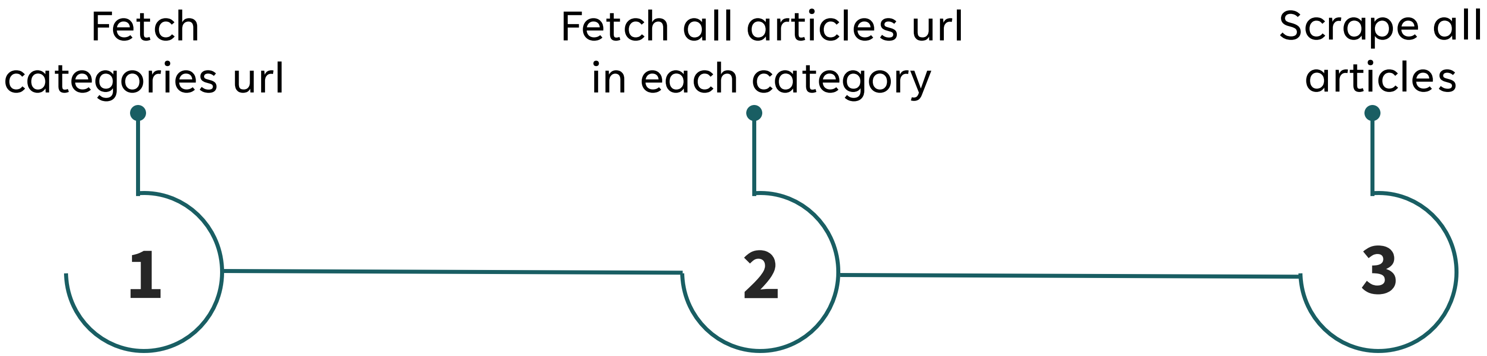


DATASET DESCRIPTION

Top commons words in each category

- FACT-CHECKING: (stat, 1408), (articol, 919), (facebook, 907), (vide, 879), (qui, 848), (legg, 823), (post, 768), (pubblic, 712), (vaccin, 672), (fals, 618)
- ATTUALITÀ: (stat, 950), (cas, 396), (prim, 386), (poi, 355), (pi, 320), (fatt, 286), (trov, 284), (anni, 266), (person, 265), (due, 259)
- CULTURA & SPETTACOLO: (stat, 592), (pi, 527), (cos, 447), (prim, 375), (anni, 338), (rai, 323), (due, 321), (pubblic, 320), (poi, 319), (perc, 317)
- EDITORIALI: (pi, 518), (prim, 508), (salvin, 495), (part, 489), (cos, 482), (fatt, 460), (stat, 451), (polit, 433), (govern, 421), (ital, 337)
- SPORT: (stat, 603), (atlet, 490), (parig, 414), (final, 394), (dop, 384), (part, 365), (prim, 349), (olimpiad, 332), (pi, 331), (gar, 297)
- AMBIENTE: (pi, 1181), (stat, 805), (climat, 533), (paes, 446), (camb, 431), (ital, 409), (nuov, 407), (esser, 376), (europe, 374), (second, 368)
- POLITICA: (stat, 770), (part, 696), (ital, 599), (pi, 591), (polit, 463), (govern, 422), (prim, 391), (president, 380), (melon, 361), (sol, 325)
- SCIENZE: (pi, 1005), (stat, 940), (ricerc, 615), (stud, 593), (vaccin, 588), (prim, 581), (esser, 504), (cas, 489), (nuov, 464), (sol, 448)
- TECNOLOGIA: (pi, 628), (stat, 598), (utent, 371), (esser, 342), (dat, 341), (nuov, 340), (piattaform, 327), (prim, 320), (propr, 314), (artificial, 304)
- LE NOSTRE STORIE: (pi, 1766), (stat, 1365), (lavor, 1181), (cos, 1043), (prim, 967), (anni, 860), (cas, 813), (sol, 773), (perc, 772), (far, 749)
- MONDO: (stat, 1124), (pi, 461), (israel, 448), (prim, 439), (russ, 403), (part, 393), (second, 381), (president, 341), (harris, 337), (dop, 331)
- ECONOMIA & LAVORO: (pi, 960), (lavor, 754), (stat, 639), (eur, 618), (ital, 583), (anni, 462), (europe, 454), (part, 445), (prim, 432), (nuov, 418)

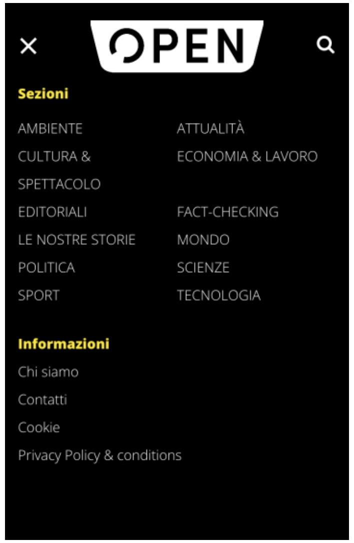
THE SCRAPER PIPELINE



1 Fetch categories url

2 Fetch all articles url in each category

3 Scrape all articles

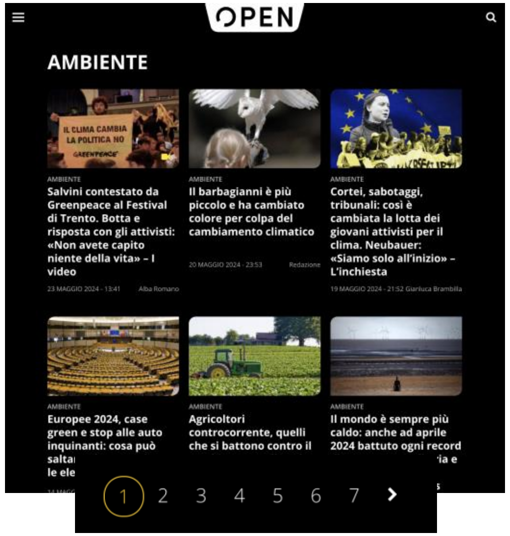


Sezioni

AMBIENTE
ATTUALITÀ
CULTURA & SPETTACOLO
ECONOMIA & LAVORO
EDITORIALI
FACT-CHECKING
LE NOSTRE STORIE
MONDO
POLITICA
SCIENZE
SPORT
TECNOLOGIA

Informazioni

Chi siamo
Contatti
Cookie
Privacy Policy & conditions



AMBIENTE

AMBIENTE
Salvini contestato da Greenpeace al Festival di Trento. Botta e risposta con gli attivisti: «Non avete capito niente della vita» – I video
23 MAGGIO 2024 - 13:41 Alba Romano

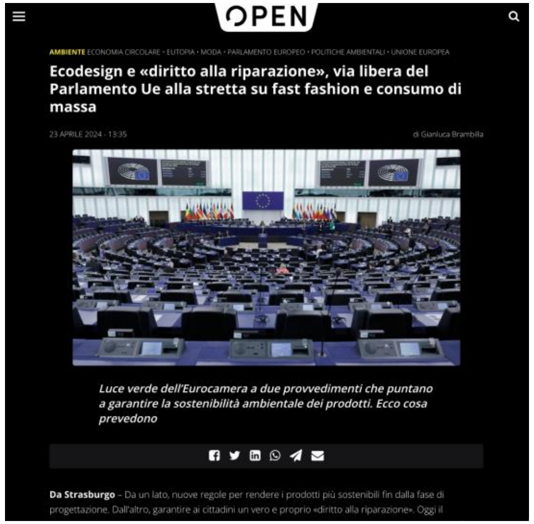
AMBIENTE
Il barbagianni è più piccolo e ha cambiato colore per colpa del cambiamento climatico
20 MAGGIO 2024 - 23:53 Redazione

AMBIENTE
Cortei, sabotaggi, tribunali: così è cambiata la lotta dei giovani attivisti per il clima. Neubauer: «Siamo solo all'inizio» – L'inchiesta
19 MAGGIO 2024 - 21:52 Gianluca Brambilla

AMBIENTE
Europee 2024, case green e stop alle auto inquinanti: cosa può saltar
le ele

AMBIENTE
Agricoltori controcorrente, quelli che si battono contro il

AMBIENTE
Il mondo è sempre più caldo: anche ad aprile 2024 battuto ogni record
ia e



AMBIENTE ECONOMIA CIRCOLARE • EUTOPIA • MODA • PARLAMENTO EUROPEO • POLITICHE AMBIENTALI • UNIONE EUROPEA

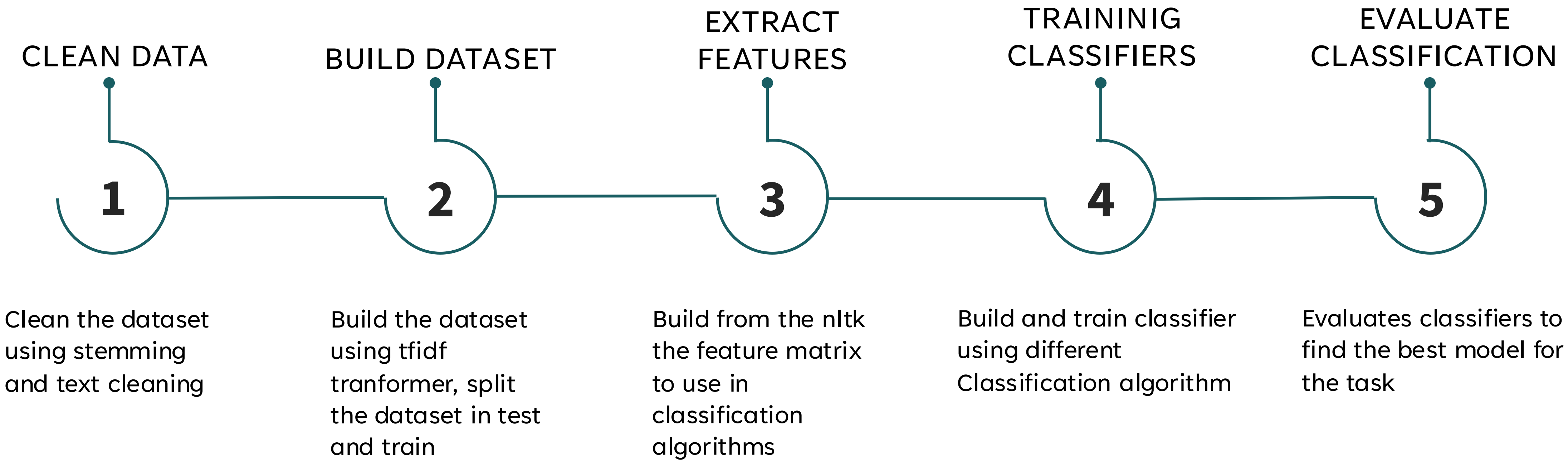
Ecodesign e «diritto alla riparazione», via libera del Parlamento Ue alla stretta su fast fashion e consumo di massa

23 APRILE 2024 - 13:35 di Gianluca Brambilla

Luce verde dell'Eurocamera a due provvedimenti che puntano a garantire la sostenibilità ambientale dei prodotti. Ecco cosa prevedono

Da Strasburgo – Da un lato, nuove regole per rendere i prodotti più sostenibili fin dalla fase di progettazione. Dall'altro, garantire ai cittadini un vero e proprio «diritto alla riparazione». Oggi il

TRAINING PIPELINE



1 CLEAN DATA

Clean the dataset using stemming and text cleaning

2 BUILD DATASET

Build the dataset using tfidf tranformer, split the dataset in test and train

3 EXTRACT FEATURES

Build from the nltk the feature matrix to use in classification algorithms

4 TRAININIG CLASSIFIERS

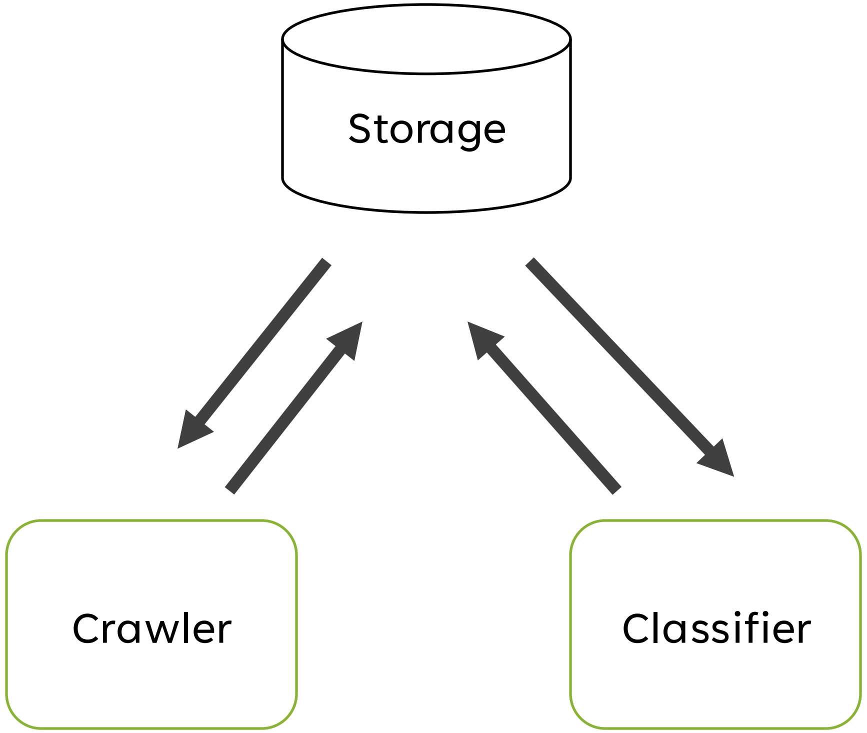
Build and train classifier using different Classification algorithm

5 EVALUATE CLASSIFICATION

Evaluates classifiers to find the best model for the task

APPLICATION STRUCTURE

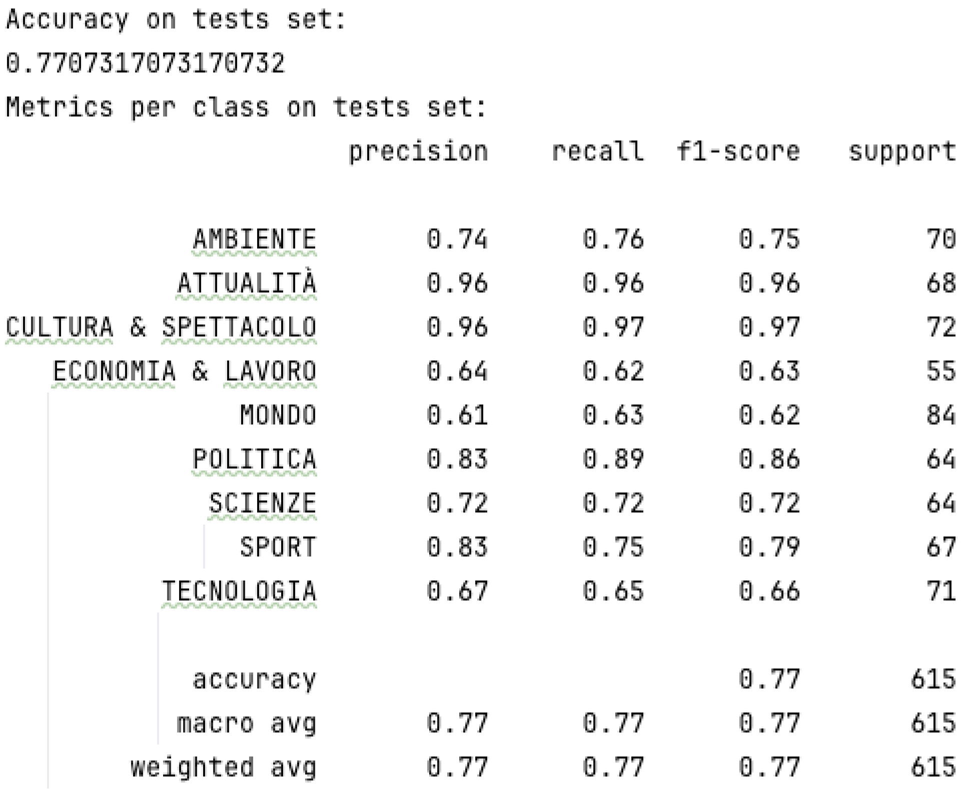
- Application is a series of python scripts divided in 2 main folders
  - Crawler
  - Classifier



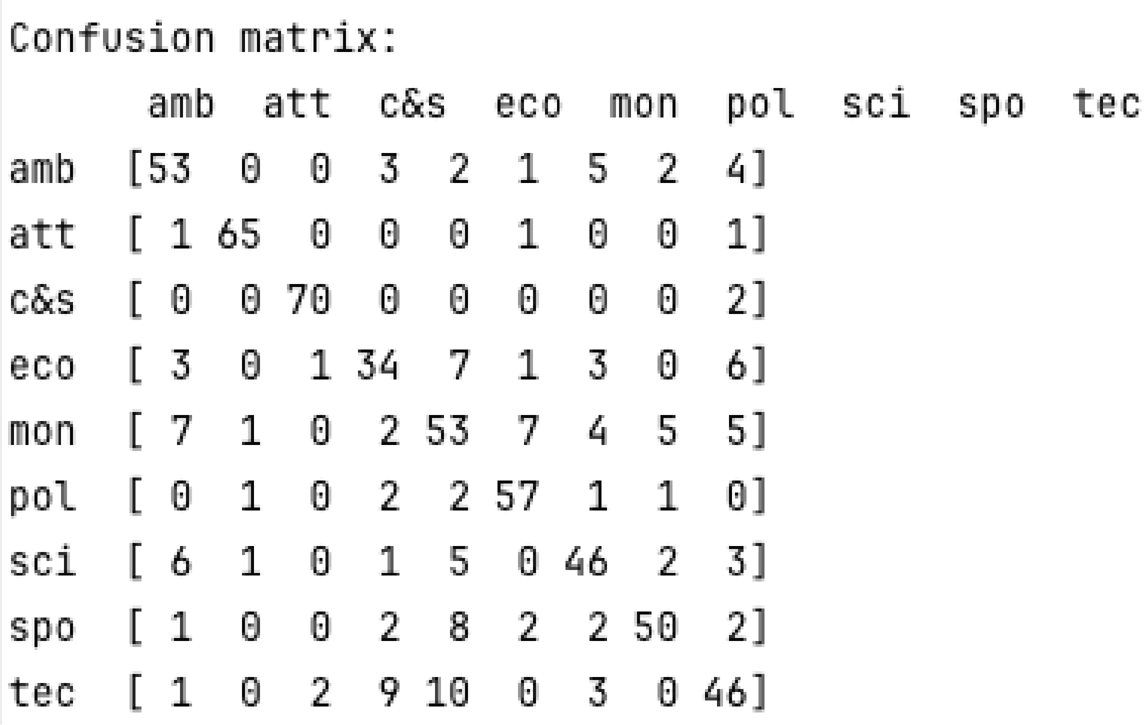
CLASSIFIERS

- Decision tree
- KNeighbors
- Multinomial Naive Bayes
- Random Forest

RESULTS - Decision Tree

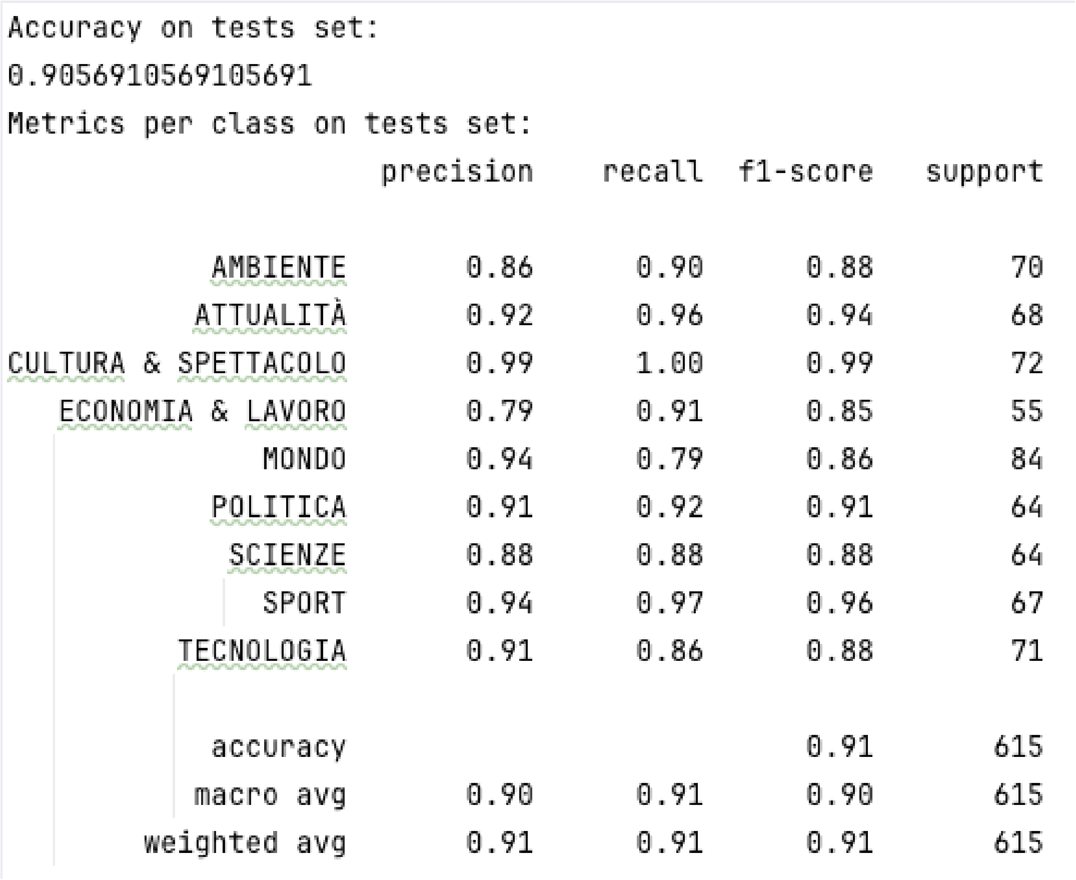


```
Accuracy on tests set:
0.7707317073170732
Metrics per class on tests set:
                      precision    recall  f1-score   support

            AMBIENTE       0.74      0.76      0.75        70
            ATTUALITÀ      0.96      0.96      0.96        68
CULTURA & SPETTACOLO       0.96      0.97      0.97        72
   ECONOMIA & LAVORO       0.64      0.62      0.63        55
               MONDO       0.61      0.63      0.62        84
            POLITICA       0.83      0.89      0.86        64
             SCIENZE       0.72      0.72      0.72        64
               SPORT       0.83      0.75      0.79        67
          TECNOLOGIA       0.67      0.65      0.66        71

            accuracy                           0.77       615
           macro avg       0.77      0.77      0.77       615
        weighted avg       0.77      0.77      0.77       615
```

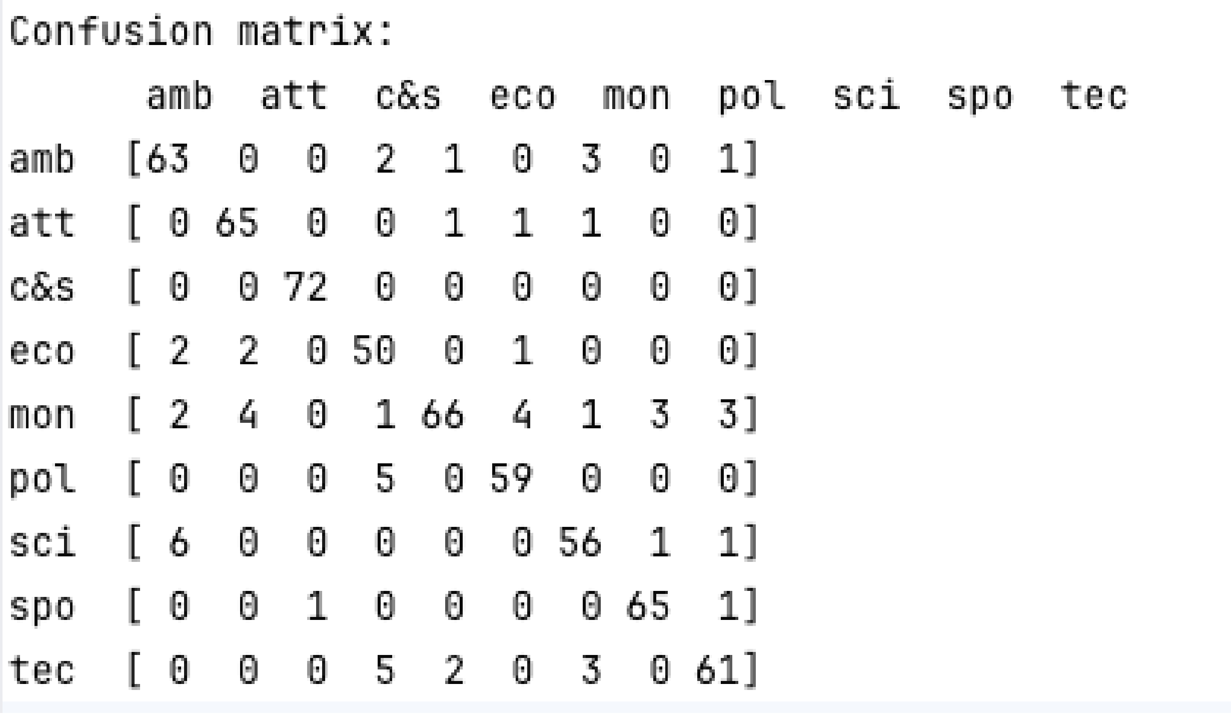


```
Confusion matrix:
      amb  att  c&s  eco  mon  pol  sci  spo  tec
amb  [53   0   0   3   2   1   5   2   4]
att  [ 1  65   0   0   0   1   0   0   1]
c&s  [ 0   0  70   0   0   0   0   0   2]
eco  [ 3   0   1  34   7   1   3   0   6]
mon  [ 7   1   0   2  53   7   4   5   5]
pol  [ 0   1   0   2   2  57   1   1   0]
sci  [ 6   1   0   1   5   0  46   2   3]
spo  [ 1   0   0   2   8   2   2  50   2]
tec  [ 1   0   2   9  10   0   3   0  46]
```

RESULTS - Random Forest

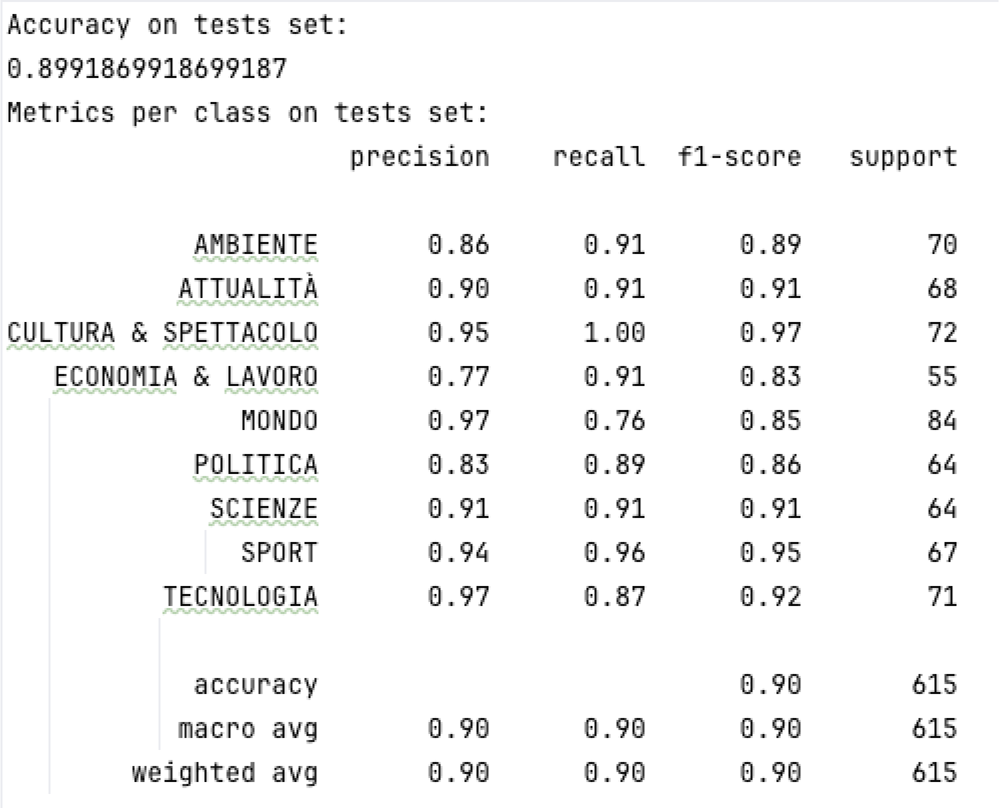


```
Accuracy on tests set:
0.9056910569105691
Metrics per class on tests set:
                                precision    recall  f1-score   support

           AMBIENTE       0.86      0.90      0.88        70
          ATTUALITÀ       0.92      0.96      0.94        68
CULTURA & SPETTACOLO       0.99      1.00      0.99        72
   ECONOMIA & LAVORO       0.79      0.91      0.85        55
               MONDO       0.94      0.79      0.86        84
            POLITICA       0.91      0.92      0.91        64
             SCIENZE       0.88      0.88      0.88        64
               SPORT       0.94      0.97      0.96        67
          TECNOLOGIA       0.91      0.86      0.88        71

            accuracy                           0.91       615
           macro avg       0.90      0.91      0.90       615
        weighted avg       0.91      0.91      0.91       615
```

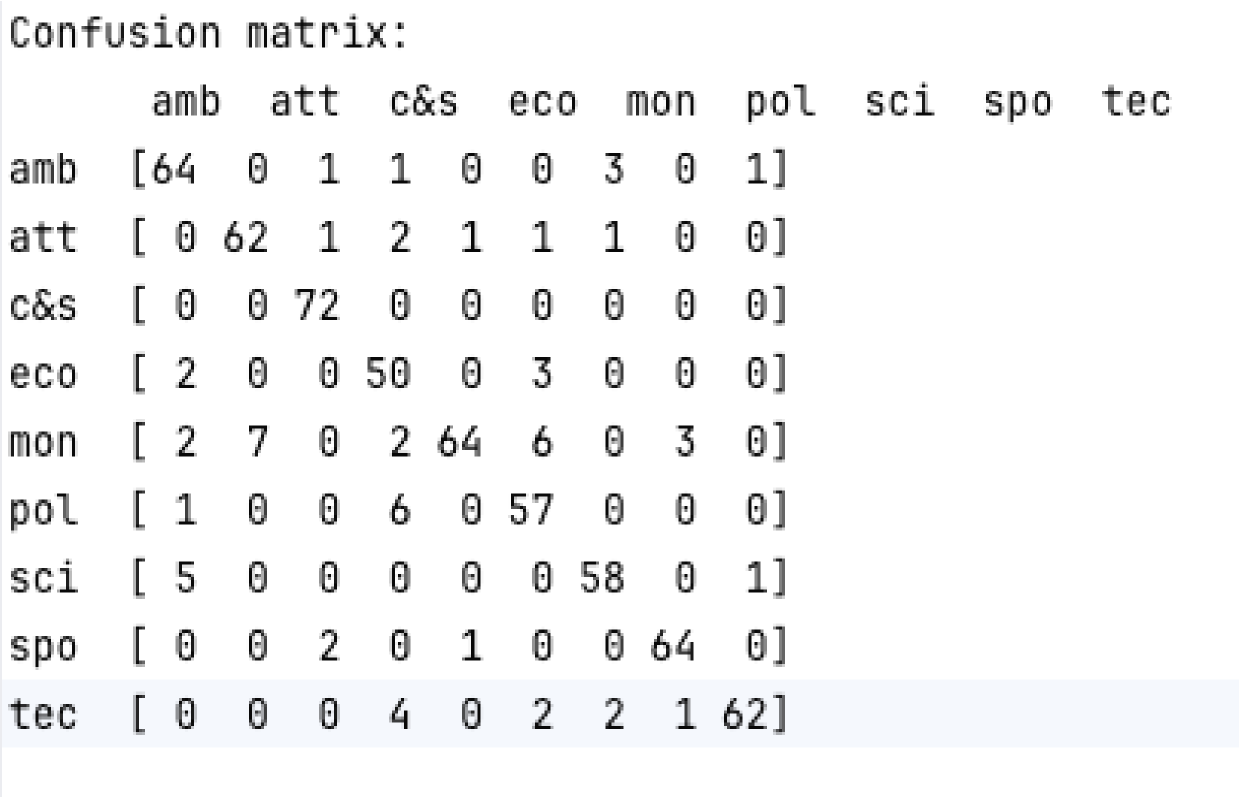


```
Confusion matrix:
      amb  att  c&s  eco  mon  pol  sci  spo  tec
amb  [63  0  0  2  1  0  3  0  1]
att  [ 0 65  0  0  1  1  1  0  0]
c&s  [ 0  0 72  0  0  0  0  0  0]
eco  [ 2  2  0 50  0  1  0  0  0]
mon  [ 2  4  0  1 66  4  1  3  3]
pol  [ 0  0  0  5  0 59  0  0  0]
sci  [ 6  0  0  0  0  0 56  1  1]
spo  [ 0  0  1  0  0  0  0 65  1]
tec  [ 0  0  0  5  2  0  3  0 61]
```

RESULTS – Naive Bayes

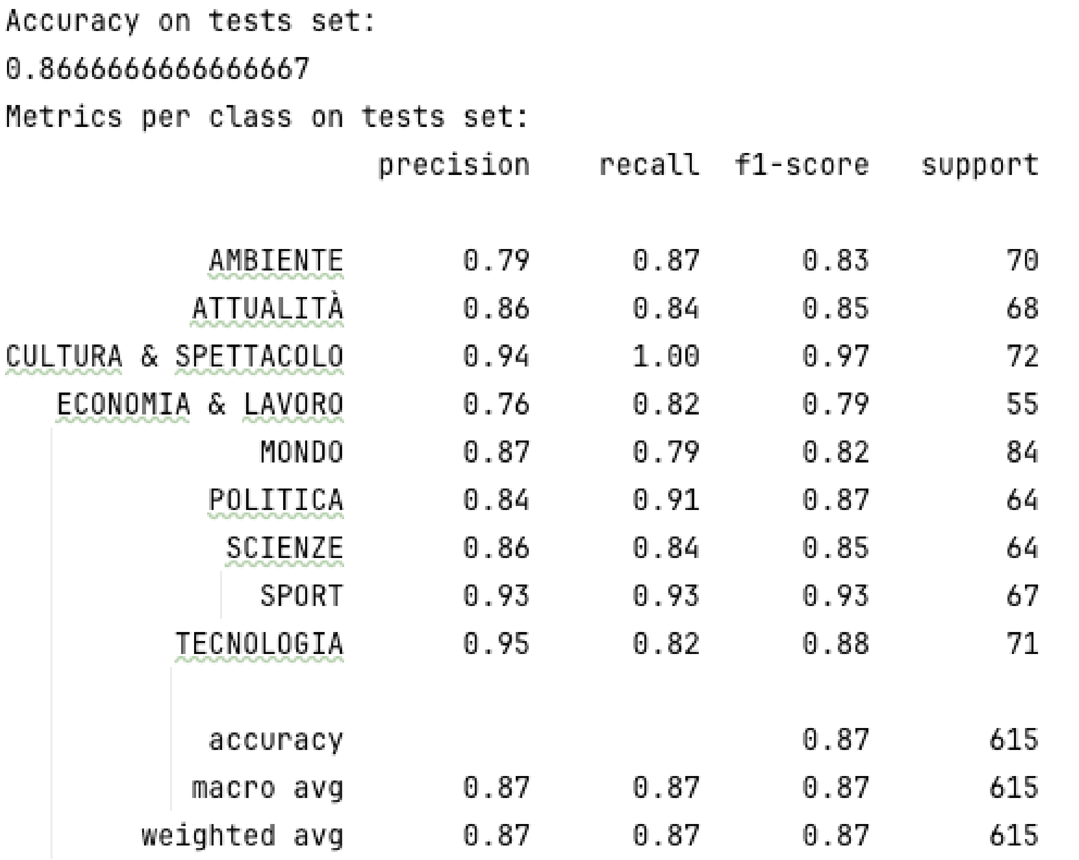


```
Accuracy on tests set:
0.8991869918699187
Metrics per class on tests set:
                     precision    recall  f1-score   support

           AMBIENTE       0.86      0.91      0.89        70
           ATTUALITÀ      0.90      0.91      0.91        68
CULTURA & SPETTACOLO      0.95      1.00      0.97        72
   ECONOMIA & LAVORO      0.77      0.91      0.83        55
               MONDO      0.97      0.76      0.85        84
            POLITICA      0.83      0.89      0.86        64
             SCIENZE      0.91      0.91      0.91        64
               SPORT      0.94      0.96      0.95        67
         TECNOLOGIA       0.97      0.87      0.92        71

            accuracy                          0.90       615
           macro avg      0.90      0.90      0.90       615
        weighted avg      0.90      0.90      0.90       615
```

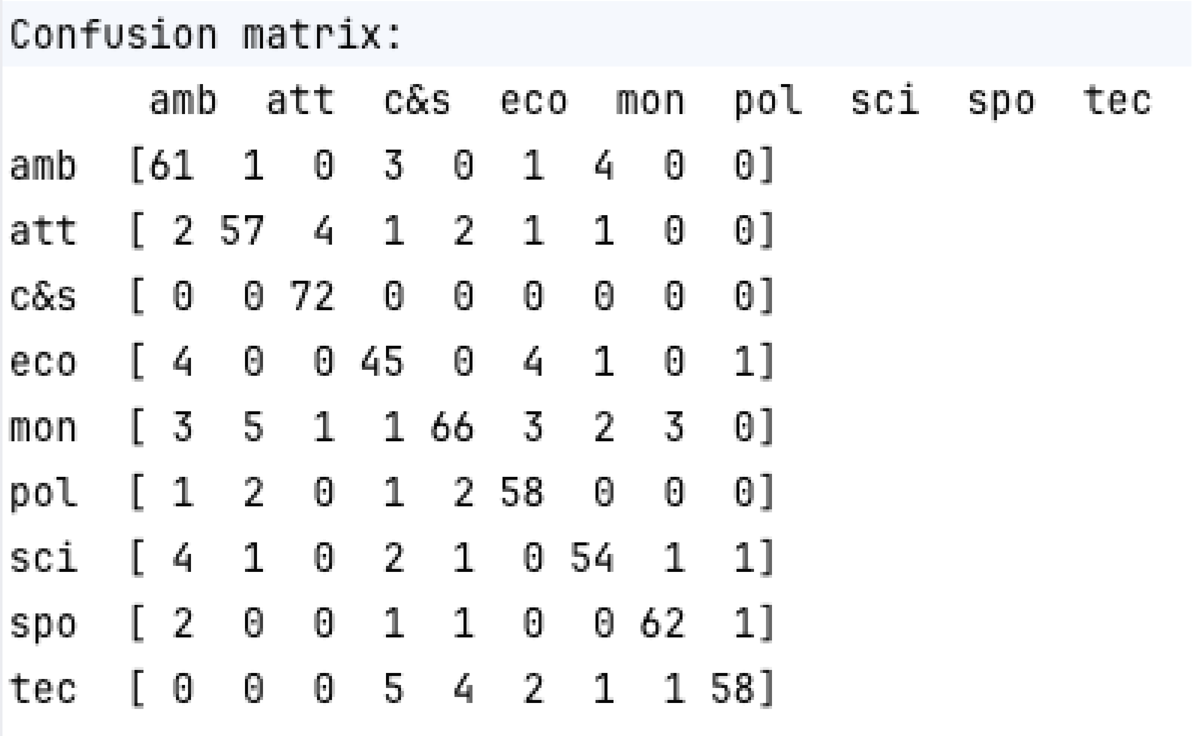


```
Confusion matrix:
       amb  att  c&s  eco  mon  pol  sci  spo  tec
amb  [64  0  1  1  0  0  3  0  1]
att  [ 0 62  1  2  1  1  1  0  0]
c&s  [ 0  0 72  0  0  0  0  0  0]
eco  [ 2  0  0 50  0  3  0  0  0]
mon  [ 2  7  0  2 64  6  0  3  0]
pol  [ 1  0  0  6  0 57  0  0  0]
sci  [ 5  0  0  0  0  0 58  0  1]
spo  [ 0  0  2  0  1  0  0 64  0]
tec  [ 0  0  0  4  0  2  2  1 62]
```

RESULTS - kNN



```
Accuracy on tests set:
0.8666666666666667
Metrics per class on tests set:
                      precision    recall  f1-score   support

            AMBIENTE       0.79      0.87      0.83        70
           ATTUALITÀ       0.86      0.84      0.85        68
CULTURA & SPETTACOLO       0.94      1.00      0.97        72
   ECONOMIA & LAVORO       0.76      0.82      0.79        55
               MONDO       0.87      0.79      0.82        84
            POLITICA       0.84      0.91      0.87        64
             SCIENZE       0.86      0.84      0.85        64
               SPORT       0.93      0.93      0.93        67
          TECNOLOGIA       0.95      0.82      0.88        71

            accuracy                           0.87       615
           macro avg       0.87      0.87      0.87       615
        weighted avg       0.87      0.87      0.87       615
```



```
Confusion matrix:
       amb  att  c&s  eco  mon  pol  sci  spo  tec
amb  [61  1  0  3  0  1  4  0  0]
att  [ 2 57  4  1  2  1  1  0  0]
c&s  [ 0  0 72  0  0  0  0  0  0]
eco  [ 4  0  0 45  0  4  1  0  1]
mon  [ 3  5  1  1 66  3  2  3  0]
pol  [ 1  2  0  1  2 58  0  0  0]
sci  [ 4  1  0  2  1  0 54  1  1]
spo  [ 2  0  0  1  1  0  0 62  1]
tec  [ 0  0  0  5  4  2  1  1 58]
```

Cross-Validation

Decision Tree Results:

- Decision Tree Cross-Validation Scores: [0.79545455, 0.78030303, 0.77272727, 0.83333333, 0.76425856, 0.74144487, 0.75285171, 0.82509506, 0.79087452, 0.78326996]
- Decision Tree Cross-Validation Mean Accuracy: 0.7839612858624265

Random Forest Results:

- Random Forest Cross-Validation Scores: [0.90151515, 0.92045455, 0.92045455, 0.92045455, 0.91254753, 0.87452471, 0.88973384, 0.92775665, 0.88212928, 0.90874525]
- Random Forest Cross-Validation Mean Accuracy: 0.9058316050236203

KNN Results:

- Cross-Validation Scores: [0.85606061, 0.82575758, 0.86742424, 0.84848485, 0.87452471, 0.79467681, 0.84410646, 0.84790875, 0.79847909, 0.8365019]
- kNN Cross-Validation Mean Accuracy: 0.8393924991358451

Naïve Bayes

- Naive Bayes Cross-Validation Scores: [0.90151515, 0.89772727, 0.90151515, 0.89015152, 0.91634981, 0.86311787, 0.87072243, 0.92395437, 0.88212928, 0.8973384]
- Naive Bayes Cross-Validation Mean Accuracy: 0.8944521258209471