**University of Pisa**

Laurea Magistrale (MSc) in Artificial Intelligence and Data Engineering

**Project**

Data Mining and Machine Learning

# Article Categorizer

Academic year 2020-2021

Lorenzo Bianchi

**Github**: https://github.com/lorebianchi98/ArticlesCategorizer.git

## Index

# Introduction

My objective is to build an application that will exploit text mining in order to automatically categorize newspapers articles with the right topic (for example sport, politics, science…).

This application could be used to categorize articles written on blogs or even post on social networks made by important accounts.

# Dataset

The dataset is built scraping articles from the Italian website Open (https://www.open.online/). This website is an online Italian free newspaper and it provides articles labelled with the right topic.

*Figure 1 Section of the online newspaper Open*

The categories in which the articles from Open are classified are the following:

- Ambiente (Environment)
- Attualità (Actuality)
- Cultura e Spettacolo (Culture and Entertainment)
- Economia e Lavoro (Economy and Work)
- Editoriali (Editorials)
- Fact-checking
- Le nostre storie (Our stories)
- Mondo (World)
- Politica (Politics)
- Scienze (Science)
- Sport
- Tecnologia (Technology)

In order to have strict categories I decided to not include in the application the category Editorials, Fact-checking and Our stories, since they contain particular articles always relatable to one of the other categories, since the purpose of this articles is not to inform readers with objective news but to release interviews or checking the truthfulness of other news. So if we use this articles to build our training set the classifier will be biased, since in this category there will be articles relatable to other categories. The category that our application will recognise will be:

- Ambiente (Environment)
- Attualità (Actuality)
- Cultura e Spettacolo (Culture and Entertainment)
- Economia e Lavoro (Economy and Work)
- Mondo (World)
- Politica (Politics)
- Scienze (Science)
- Sport
- Tecnologia (Technology)

We will construct our dataset scraping 300 articles for each category (2700 in total). The 80% of the dataset will compose our training and the remaining 20% will compose the test set, and it will be used to evaluate the effectiveness of the application.

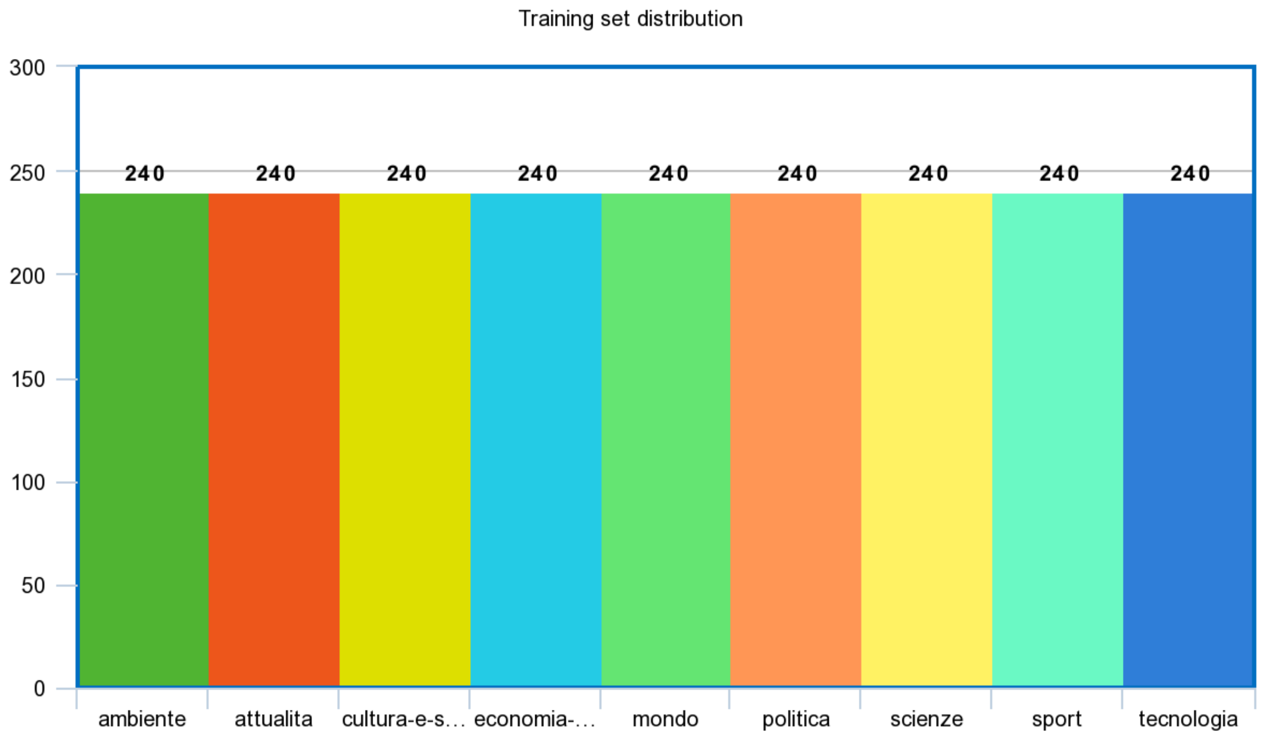


*Figure 2 Distribution of the training set*

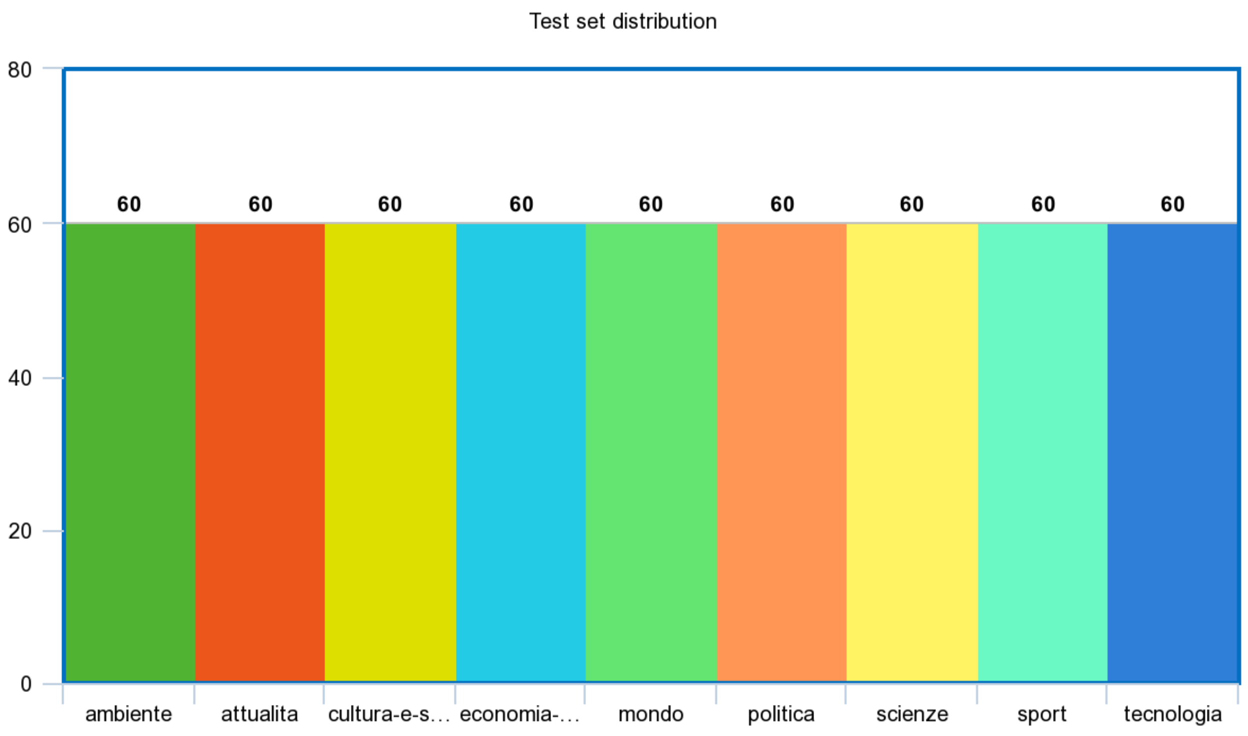


*Figure 3 Distribution of the test set*

# The Application

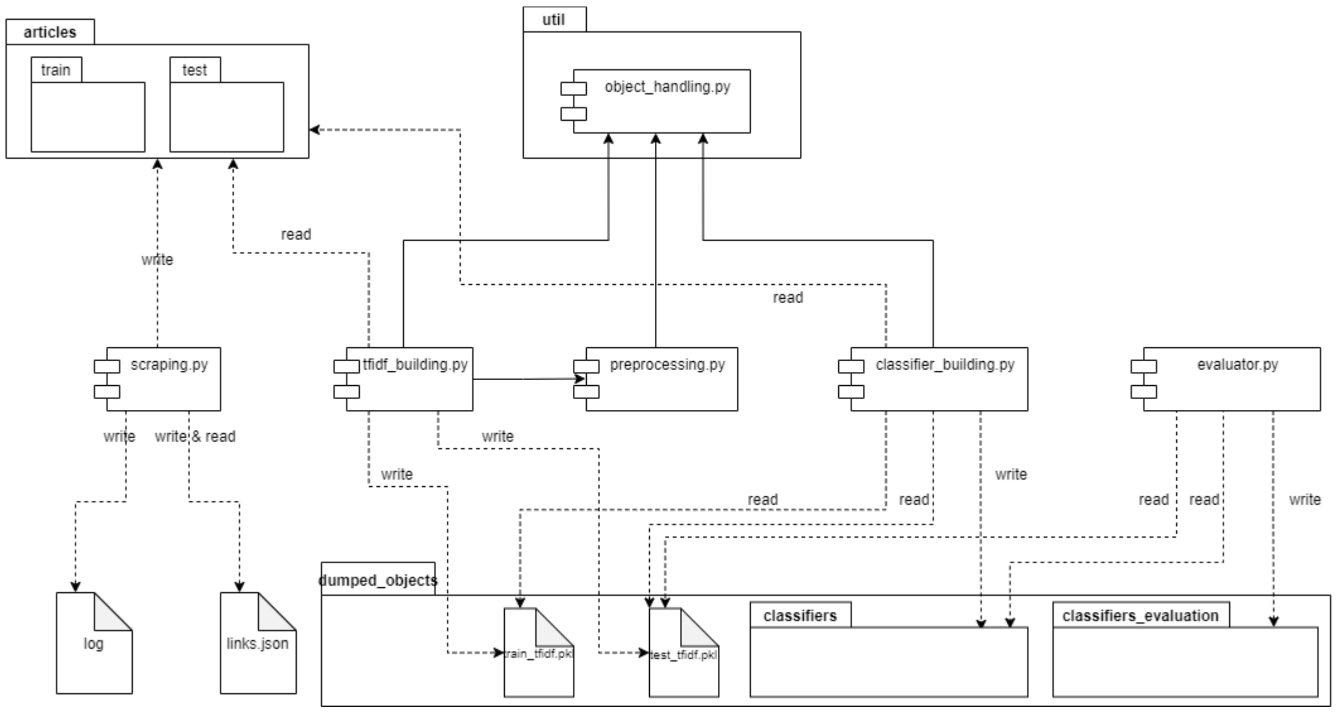
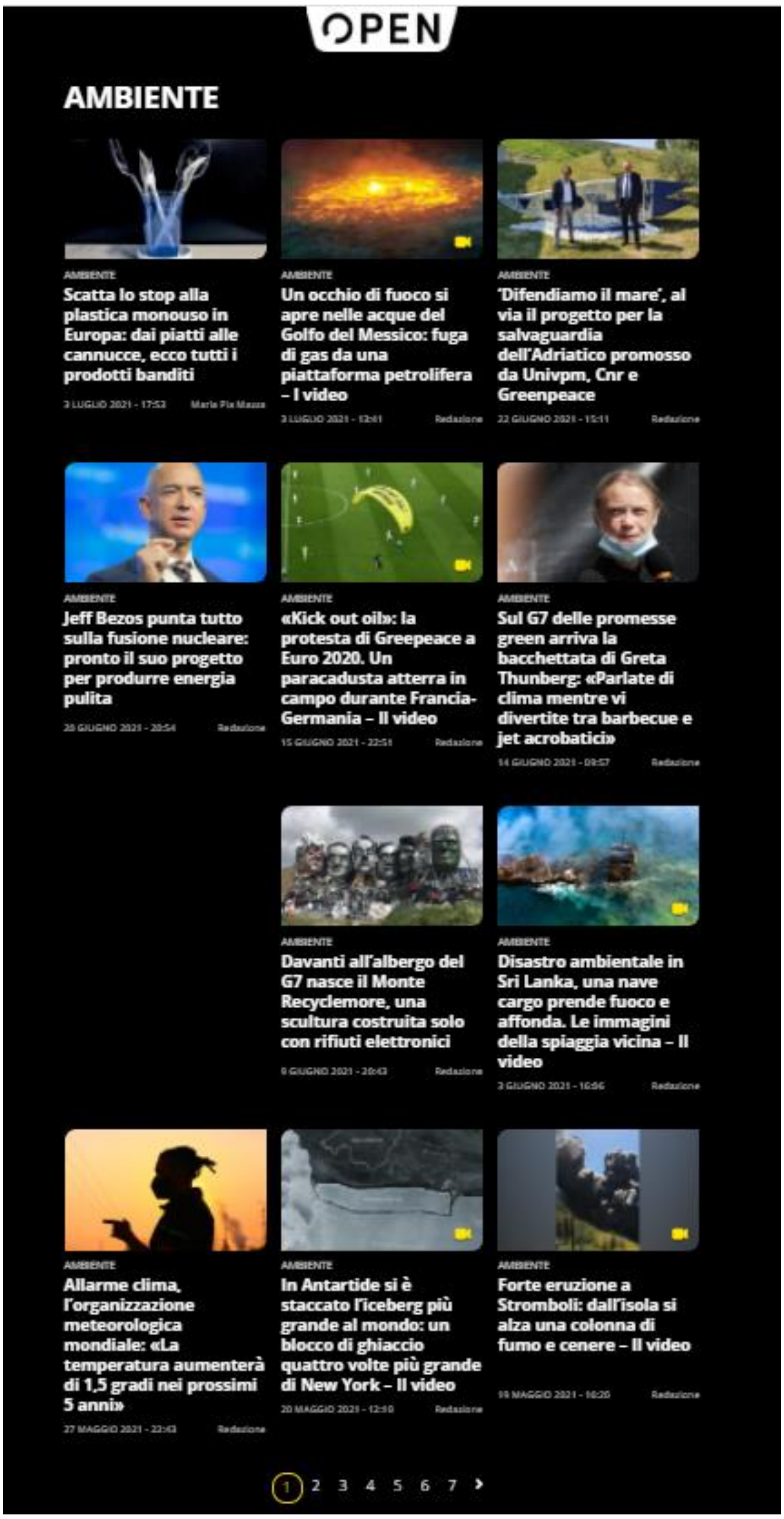


*Figure 4 UML diagram of the application*

## Scraping

The scraping phase is divided in two phase: the link harvesting and the article scraping.

The link harvesting is executed using the function linkScraper(url, numArticles), which collects the links from the page of the sections of Open (the base url is passed to the function). The number of link scraped for each categories is defined by numArticles. The function iterates the page that contains the article of each section, each page contains 12 articles, so for each category the script needs to pass over 25 pages. The links are stored in a human readable format (json) on a file links.json.


OPEN

AMBIENTE

AMBIENTE
Scatta lo stop alla plastica monouso in Europa: dai piatti alle cannucce, ecco tutti i prodotti banditi
3 LUGLIO 2021 - 17:53 Maria Pia Mazza

AMBIENTE
Un occhio di fuoco si apre nelle acque del Golfo del Messico: fuga di gas da una piattaforma petrolifera – I video
3 LUGLIO 2021 - 13:41 Redazione

AMBIENTE
'Difendiamo il mare', al via il progetto per la salvaguardia dell'Adriatico promosso da Univpm, Cnr e Greenpeace
22 GIUGNO 2021 - 15:11 Redazione

AMBIENTE
Jeff Bezos punta tutto sulla fusione nucleare: pronto il suo progetto per produrre energia pulita
20 GIUGNO 2021 - 20:54 Redazione

AMBIENTE
«Kick out oil»: la protesta di Greepeace a Euro 2020. Un paracadusta atterra in campo durante Francia-Germania – Il video
15 GIUGNO 2021 - 22:51 Redazione

AMBIENTE
Sul G7 delle promesse green arriva la bacchettata di Greta Thunberg: «Parlate di clima mentre vi divertite tra barbecue e jet acrobatici»
14 GIUGNO 2021 - 09:57 Redazione

AMBIENTE
Davanti all'albergo del G7 nasce il Monte Recyclemore, una scultura costruita solo con rifiuti elettronici
9 GIUGNO 2021 - 20:43 Redazione

AMBIENTE
Disastro ambientale in Sri Lanka, una nave cargo prende fuoco e affonda. Le immagini della spiaggia vicina – Il video
3 GIUGNO 2021 - 16:56 Redazione

AMBIENTE
Allarme clima, l'organizzazione meteorologica mondiale: «La temperatura aumenterà di 1,5 gradi nei prossimi 5 anni»
27 MAGGIO 2021 - 22:43 Redazione

AMBIENTE
In Antartide si è staccato l'iceberg più grande al mondo: un blocco di ghiaccio quattro volte più grande di New York – Il video
20 MAGGIO 2021 - 12:10 Redazione

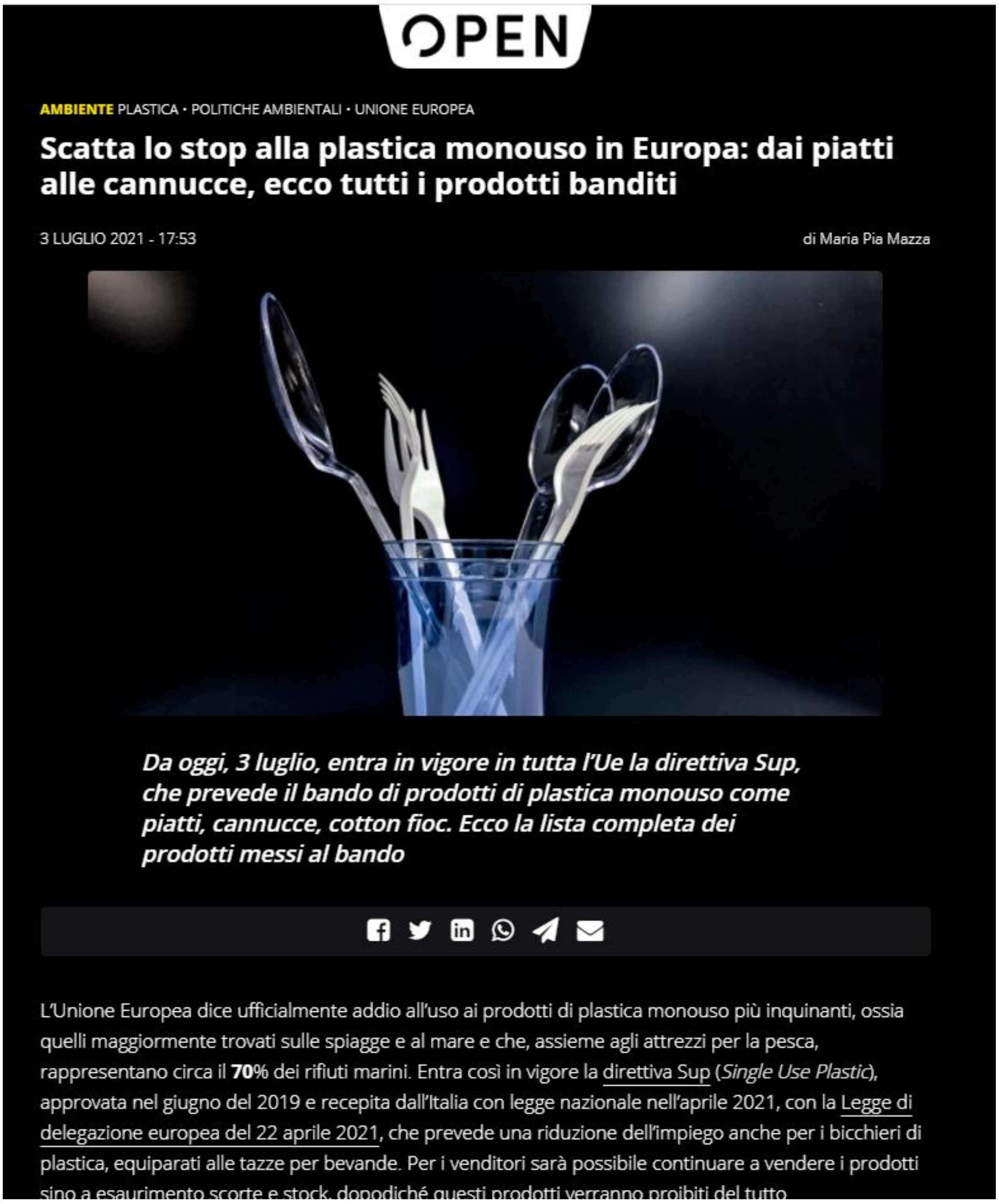
AMBIENTE
Forte eruzione a Stromboli: dall'isola si alza una colonna di fumo e cenere – Il video
19 MAGGIO 2021 - 10:20 Redazione

1 2 3 4 5 6 7 ›

*Figure 5 Page which contains the list of articles for the environment section*

The article scraping phase is executed using the function getAllArticles(firstToScrape), that start scraping articles in order from the firstToScrape (this parameter is useful to complete the scraping in different times, giving the opportunity of restarting the scraping from the last article scraped). This function reads the links.json file, and get the text of each articles using the extractText(url)

function. An article has a title (h1 tag), subtitles (h2 tags) and paragraphs (p tags). We decided to consider all this section in the text of the article, and all this tag texts are obtained using the BeautifulSoup library. The 80% of the articles are stored in articles/train/CATEGORY, the remaining 20% on articles/test/CATEGORY.
If a problem occurs during the scraping of one article, the link which caused the failure is stored in a log file, and it needs to be rescraped.



**AMBIENTE** PLASTICA • POLITICHE AMBIENTALI • UNIONE EUROPEA

**Scatta lo stop alla plastica monouso in Europa: dai piatti alle cannucce, ecco tutti i prodotti banditi**

3 LUGLIO 2021 - 17:53 di Maria Pia Mazza

***Da oggi, 3 luglio, entra in vigore in tutta l'Ue la direttiva Sup, che prevede il bando di prodotti di plastica monouso come piatti, cannucce, cotton fioc. Ecco la lista completa dei prodotti messi al bando***

L'Unione Europea dice ufficialmente addio all'uso ai prodotti di plastica monouso più inquinanti, ossia quelli maggiormente trovati sulle spiagge e al mare e che, assieme agli attrezzi per la pesca, rappresentano circa il **70%** dei rifiuti marini. Entra così in vigore la direttiva Sup (*Single Use Plastic*), approvata nel giugno del 2019 e recepita dall'Italia con legge nazionale nell'aprile 2021, con la Legge di delegazione europea del 22 aprile 2021, che prevede una riduzione dell'impiego anche per i bicchieri di plastica, equiparati alle tazze per bevande. Per i venditori sarà possibile continuare a vendere i prodotti sino a esaurimento scorte e stock, dopodiché questi prodotti verranno proibiti del tutto

Figure 6 An example of article

## Preprocessing

The preprocessing functions are written in preprocessing.py. The most important function of the module is preprocess_text(doc) that will be the analyzer function used by the CountVectorizer. This function will remove the alphanumeric character from the text, using nltk.corpus.stopwors will remove all the Italian stop words and all the words are stemmed using the ItalianStemmer provided by nltk.stem.snowball. Passing this function as analyzer parameter for the CountVectorizer, will let the CountVectorizer to tokenize the text and to perform the cleaning, the stemming and the stop word filtering.

This module contains also the function decodeData(list): some articles scraped from Open needs to be decoded, others no. This function will return a list of document all decoded and ready to be used.

## Classifier Building

The file tfidf_building.py takes as input the training set and the test set and gives as output two files containing the term frequency–inverse document frequency of the training set and of the test set (train_tfidf.pkl and test_tfidf.pkl).

Then the classifiers are built executing the script classifier_building.py that creates the following classifiers:

- Multinomial Naïve Bayes classifier
- Support Vector Machines classifier
- Decision Tree classifier
- Random Forest classifier
- K-Neighbors classifier

The k for the K-Neighbors classifier is chosen trying the classifier on the test set with all the k between 1 and 100, and the k which leds to the best accuracy is kept (the tests shown that the best parameter is k = 27).

Then the classifiers are fitted on the training set and are stored on a permanent file.

## Evaluation

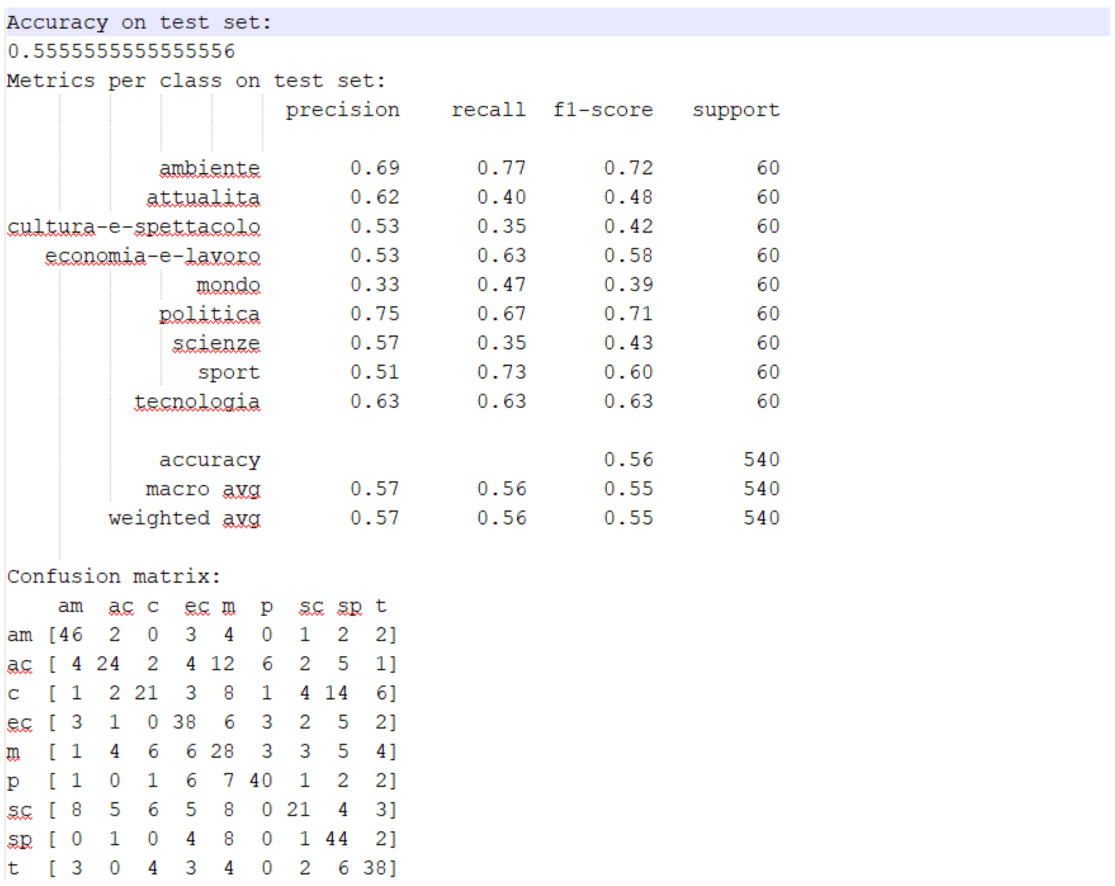
The classifiers are evaluated using the evaluator.py script, that will test all the classifiers against the test set, and it will save the results on files.

# Results

The results obtained from the evaluation phase are reported in the following pictures. The legend to read the confusion matrix is the following:

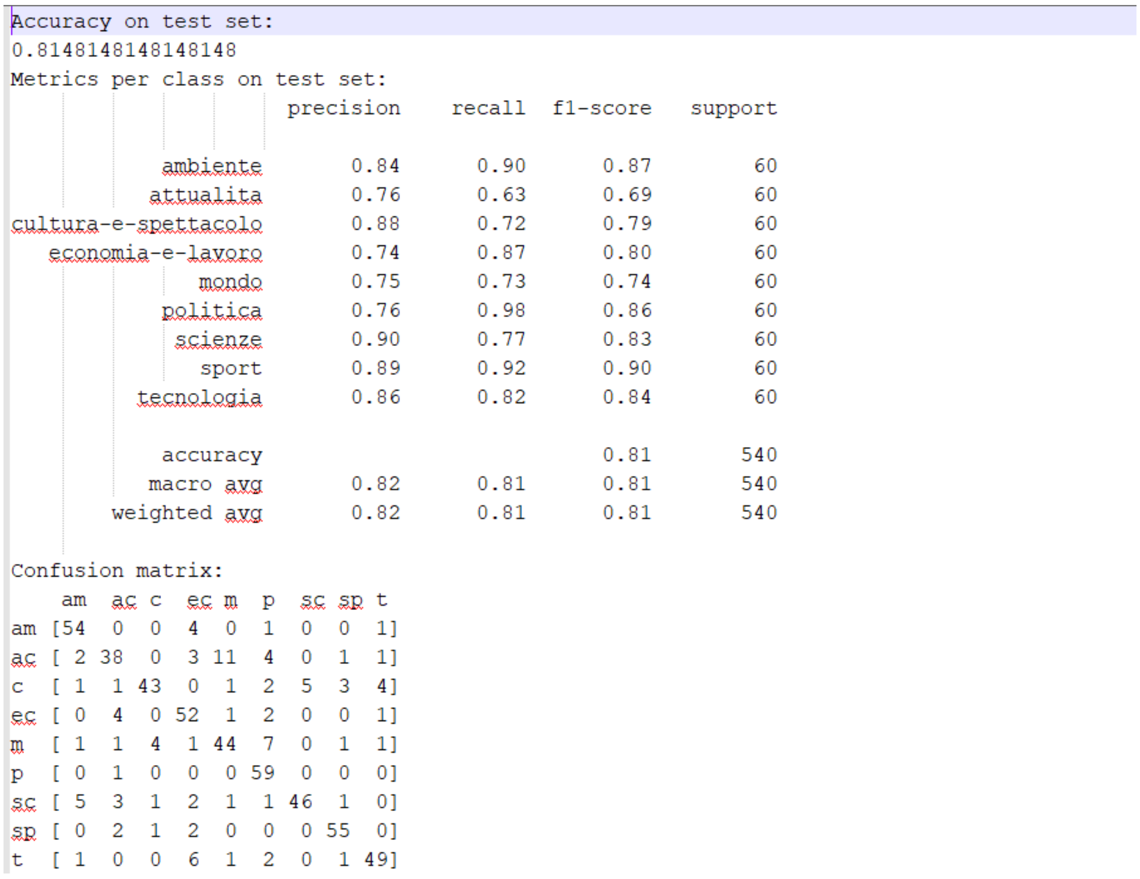
- am = Ambiente (Environment)

- ac = Attualità (Actuality)
- c = Cultura e Spettacolo (Culture and Entertainment)
- ec = Economia e Lavoro (Economy and Work)
- m = Mondo (world)
- p = Politica (Politics)
- sc = Scienze (Science)
- sp = Sport
- t = Tecnologia (Technology)



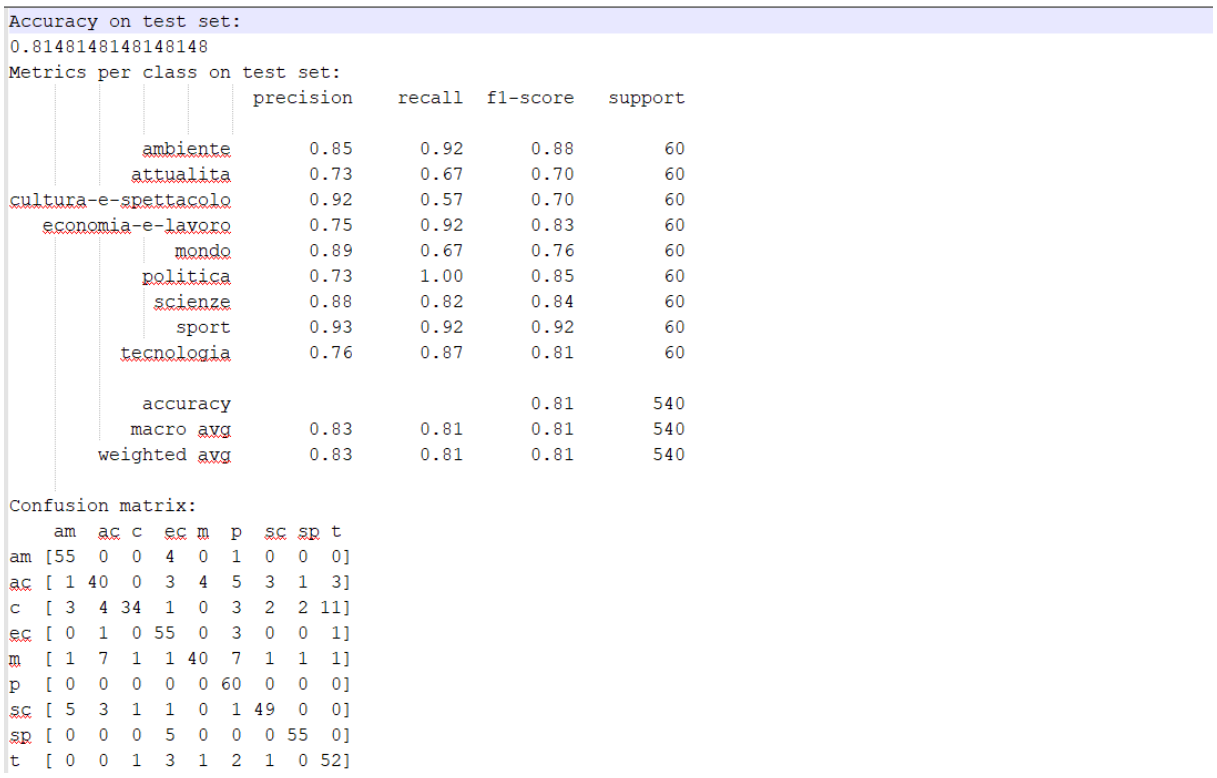
```
Accuracy on test set:
0.5555555555555556
Metrics per class on test set:
                      precision    recall  f1-score   support

           ambiente       0.69      0.77      0.72        60
          attualita       0.62      0.40      0.48        60
cultura-e-spettacolo       0.53      0.35      0.42        60
   economia-e-lavoro       0.53      0.63      0.58        60
              mondo       0.33      0.47      0.39        60
           politica       0.75      0.67      0.71        60
            scienze       0.57      0.35      0.43        60
              sport       0.51      0.73      0.60        60
         tecnologia       0.63      0.63      0.63        60

           accuracy                           0.56       540
          macro avg       0.57      0.56      0.55       540
       weighted avg       0.57      0.56      0.55       540

Confusion matrix:
    am  ac  c  ec  m   p  sc  sp  t
am  [46  2  0  3   4  0   1   2   2]
ac  [ 4 24  2  4  12  6   2   5   1]
c   [ 1  2 21  3   8  1   4  14   6]
ec  [ 3  1  0 38   6  3   2   5   2]
m   [ 1  4  6  6  28  3   3   5   4]
p   [ 1  0  1  6   7 40   1   2   2]
sc  [ 8  5  6  5   8  0  21   4   3]
sp  [ 0  1  0  4   8  0   1  44   2]
t   [ 3  0  4  3   4  0   2   6  38]
```

*Figure 7 Decision Tree results*



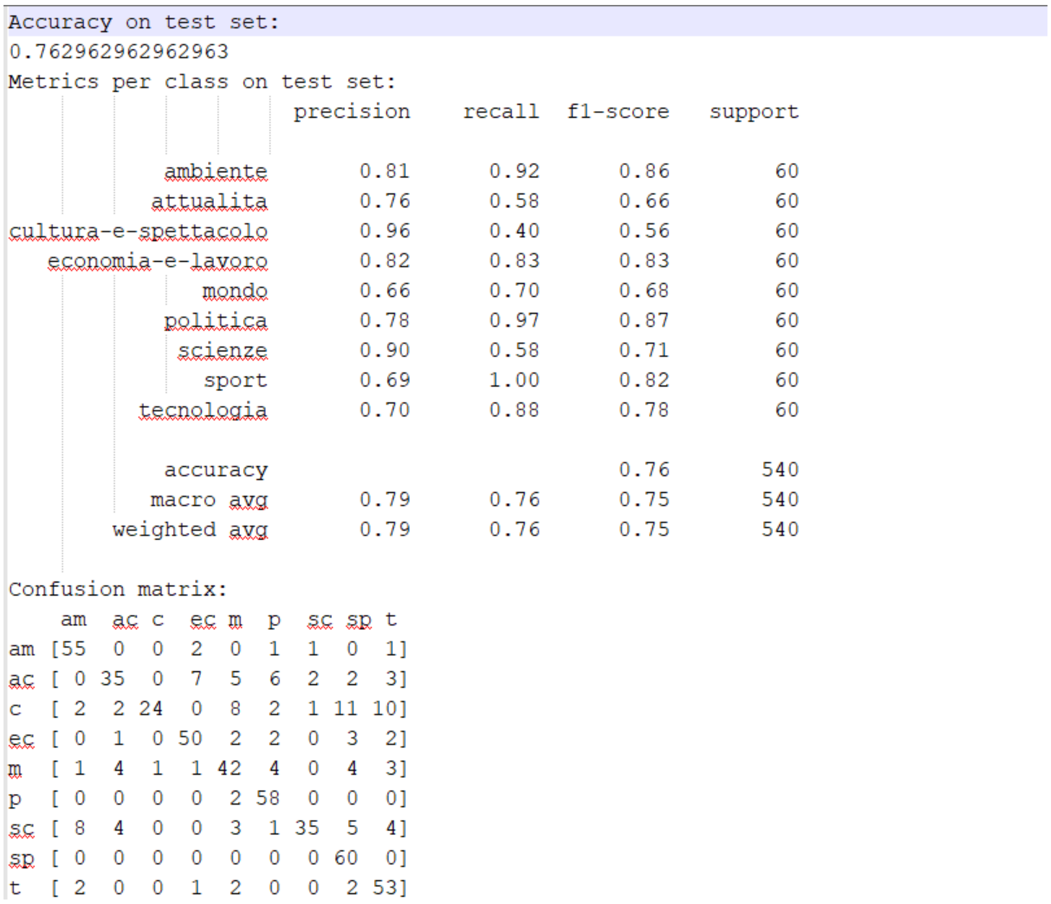
```
Accuracy on test set:
0.8148148148148148
Metrics per class on test set:
                      precision    recall  f1-score   support

            ambiente       0.84      0.90      0.87        60
           attualita       0.76      0.63      0.69        60
cultura-e-spettacolo       0.88      0.72      0.79        60
   economia-e-lavoro       0.74      0.87      0.80        60
               mondo       0.75      0.73      0.74        60
            politica       0.76      0.98      0.86        60
             scienze       0.90      0.77      0.83        60
               sport       0.89      0.92      0.90        60
           tecnologia      0.86      0.82      0.84        60

            accuracy                           0.81       540
           macro avg       0.82      0.81      0.81       540
        weighted avg       0.82      0.81      0.81       540

Confusion matrix:
    am  ac  c  ec  m   p  sc  sp  t
am  [54  0  0  4   0  1   0   0  1]
ac  [ 2 38  0  3  11  4   0   1  1]
c   [ 1  1 43  0   1  2   5   3  4]
ec  [ 0  4  0 52   1  2   0   0  1]
m   [ 1  1  4  1  44  7   0   1  1]
p   [ 0  1  0  0   0 59   0   0  0]
sc  [ 5  3  1  2   1  1  46   1  0]
sp  [ 0  2  1  2   0  0   0  55  0]
t   [ 1  0  0  6   1  2   0   1 49]
```

*Figure 8 KNeighbors results*



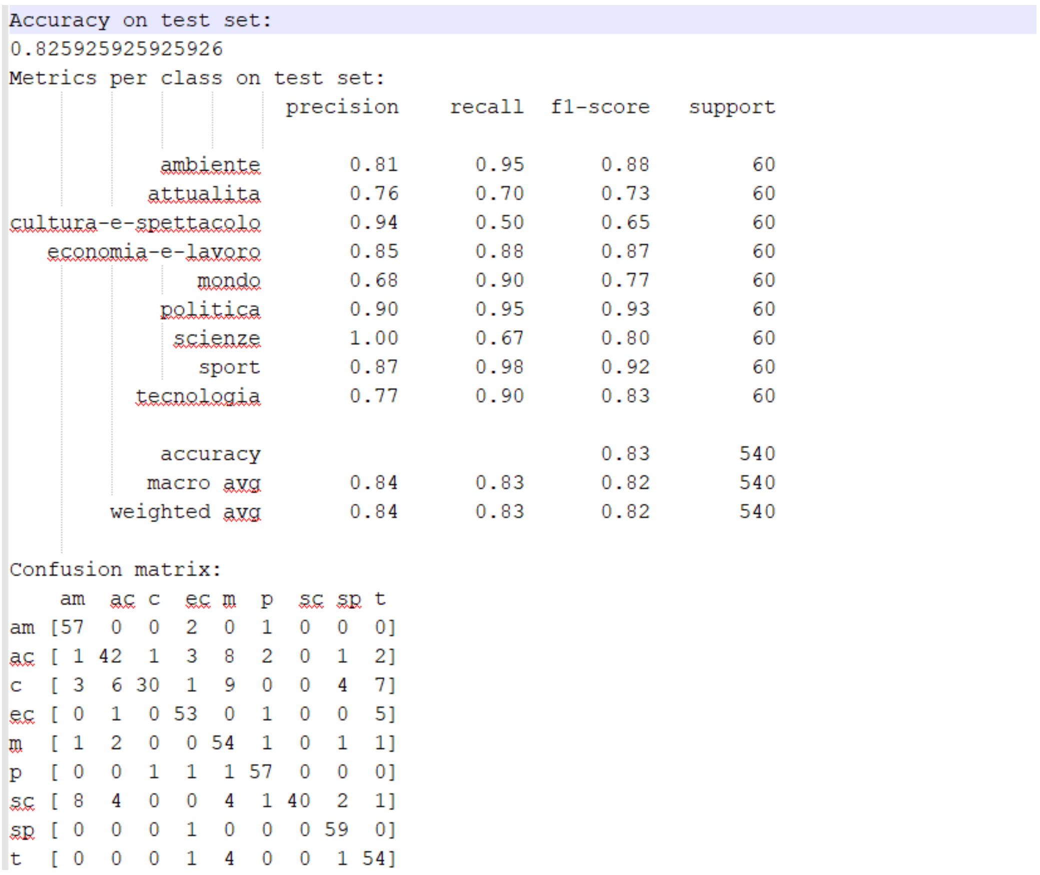
```
Accuracy on test set:
0.8148148148148148
Metrics per class on test set:
                      precision    recall  f1-score   support

            ambiente       0.85      0.92      0.88        60
           attualita       0.73      0.67      0.70        60
cultura-e-spettacolo       0.92      0.57      0.70        60
   economia-e-lavoro       0.75      0.92      0.83        60
               mondo       0.89      0.67      0.76        60
            politica       0.73      1.00      0.85        60
             scienze       0.88      0.82      0.84        60
               sport       0.93      0.92      0.92        60
           tecnologia      0.76      0.87      0.81        60

            accuracy                           0.81       540
           macro avg       0.83      0.81      0.81       540
        weighted avg       0.83      0.81      0.81       540

Confusion matrix:
    am  ac  c  ec  m   p  sc  sp  t
am  [55  0  0  4   0  1   0   0  0]
ac  [ 1 40  0  3   4  5   3   1  3]
c   [ 3  4 34  1   0  3   2   2 11]
ec  [ 0  1  0 55   0  3   0   0  1]
m   [ 1  7  1  1  40  7   1   1  1]
p   [ 0  0  0  0   0 60   0   0  0]
sc  [ 5  3  1  1   0  1  49   0  0]
sp  [ 0  0  0  5   0  0   0  55  0]
t   [ 0  0  1  3   1  2   1   0 52]
```

*Figure 9 MultinomiaNB classifier*



```
Accuracy on test set:
0.762962962962963
Metrics per class on test set:
                      precision    recall  f1-score   support

             ambiente      0.81      0.92      0.86        60
            attualita      0.76      0.58      0.66        60
cultura-e-spettacolo      0.96      0.40      0.56        60
   economia-e-lavoro      0.82      0.83      0.83        60
                mondo      0.66      0.70      0.68        60
             politica      0.78      0.97      0.87        60
              scienze      0.90      0.58      0.71        60
                sport      0.69      1.00      0.82        60
           tecnologia      0.70      0.88      0.78        60

             accuracy                          0.76       540
            macro avg      0.79      0.76      0.75       540
         weighted avg      0.79      0.76      0.75       540

Confusion matrix:
    am  ac  c  ec  m   p  sc  sp  t
am  [55   0   0   2   0   1   1   0   1]
ac  [ 0  35   0   7   5   6   2   2   3]
c   [ 2   2  24   0   8   2   1  11  10]
ec  [ 0   1   0  50   2   2   0   3   2]
m   [ 1   4   1   1  42   4   0   4   3]
p   [ 0   0   0   0   2  58   0   0   0]
sc  [ 8   4   0   0   3   1  35   5   4]
sp  [ 0   0   0   0   0   0   0  60   0]
t   [ 2   0   0   1   2   0   0   2  53]
```

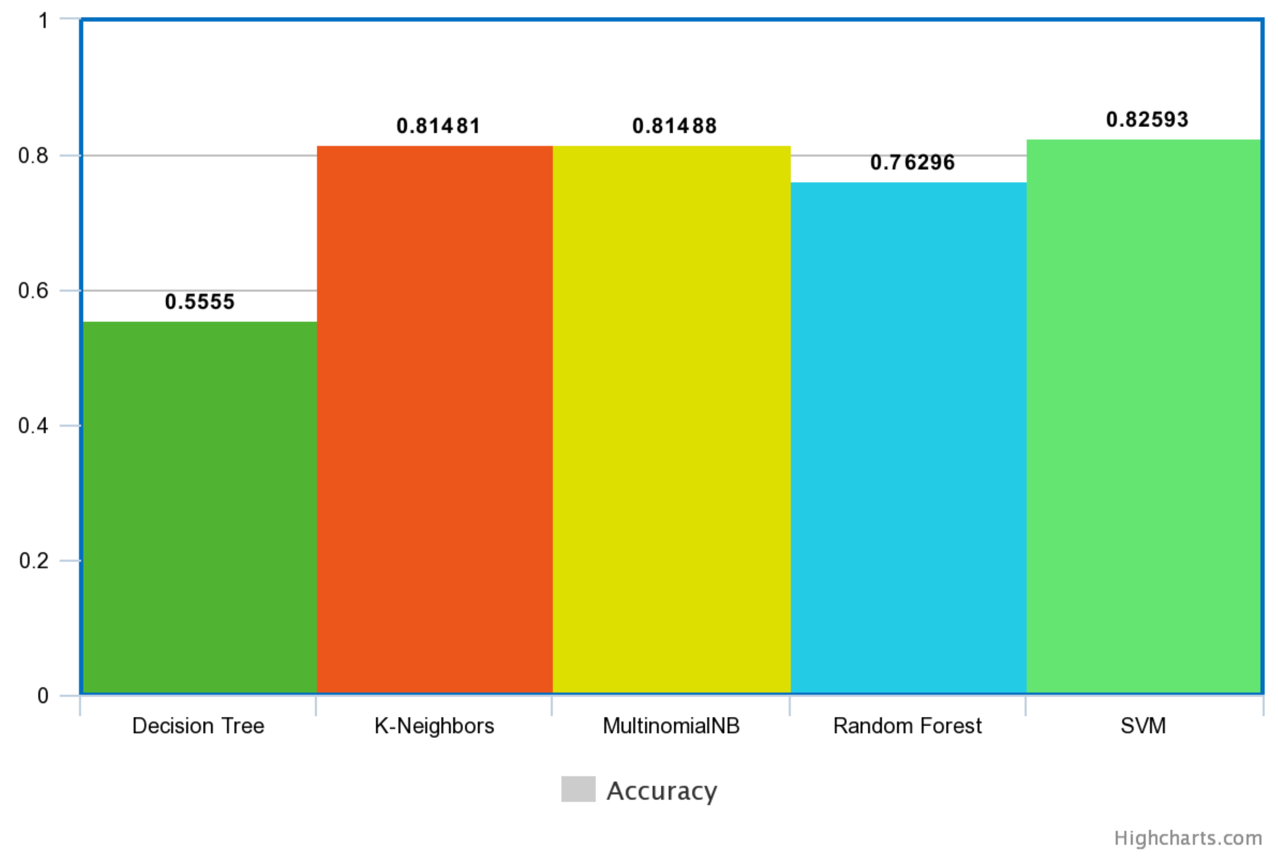
*Figure 10 Random Forest classifier*



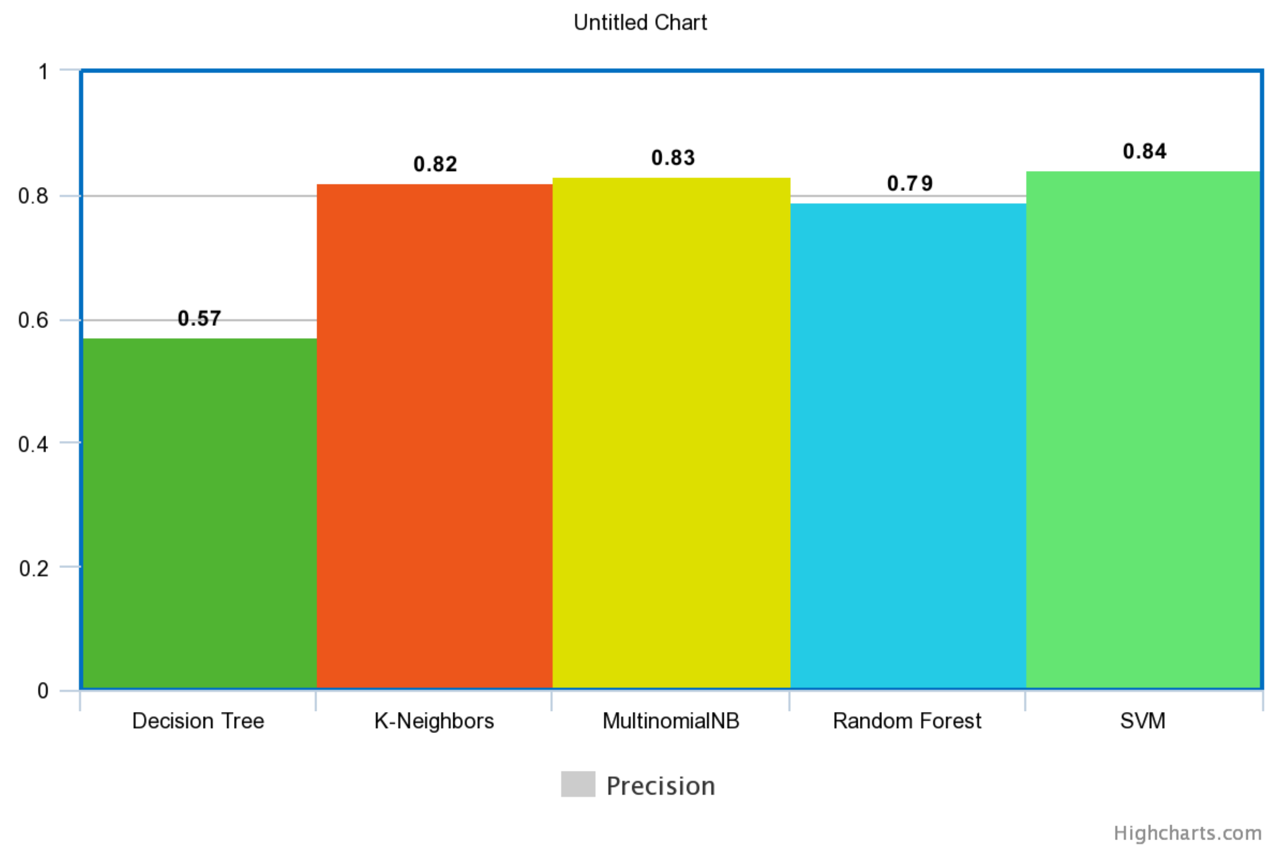
```
Accuracy on test set:
0.825925925925926
Metrics per class on test set:
                      precision    recall  f1-score   support

             ambiente      0.81      0.95      0.88        60
            attualita      0.76      0.70      0.73        60
cultura-e-spettacolo      0.94      0.50      0.65        60
   economia-e-lavoro      0.85      0.88      0.87        60
                mondo      0.68      0.90      0.77        60
             politica      0.90      0.95      0.93        60
              scienze      1.00      0.67      0.80        60
                sport      0.87      0.98      0.92        60
           tecnologia      0.77      0.90      0.83        60

             accuracy                          0.83       540
            macro avg      0.84      0.83      0.82       540
         weighted avg      0.84      0.83      0.82       540

Confusion matrix:
    am  ac  c  ec  m   p  sc  sp  t
am  [57   0   0   2   0   1   0   0   0]
ac  [ 1  42   1   3   8   2   0   1   2]
c   [ 3   6  30   1   9   0   0   4   7]
ec  [ 0   1   0  53   0   1   0   0   5]
m   [ 1   2   0   0  54   1   0   1   1]
p   [ 0   0   1   1   1  57   0   0   0]
sc  [ 8   4   0   0   4   1  40   2   1]
sp  [ 0   0   0   1   0   0   0  59   0]
t   [ 0   0   0   1   4   0   0   1  54]
```

*Figure 11 Support Vector Machines classifier*

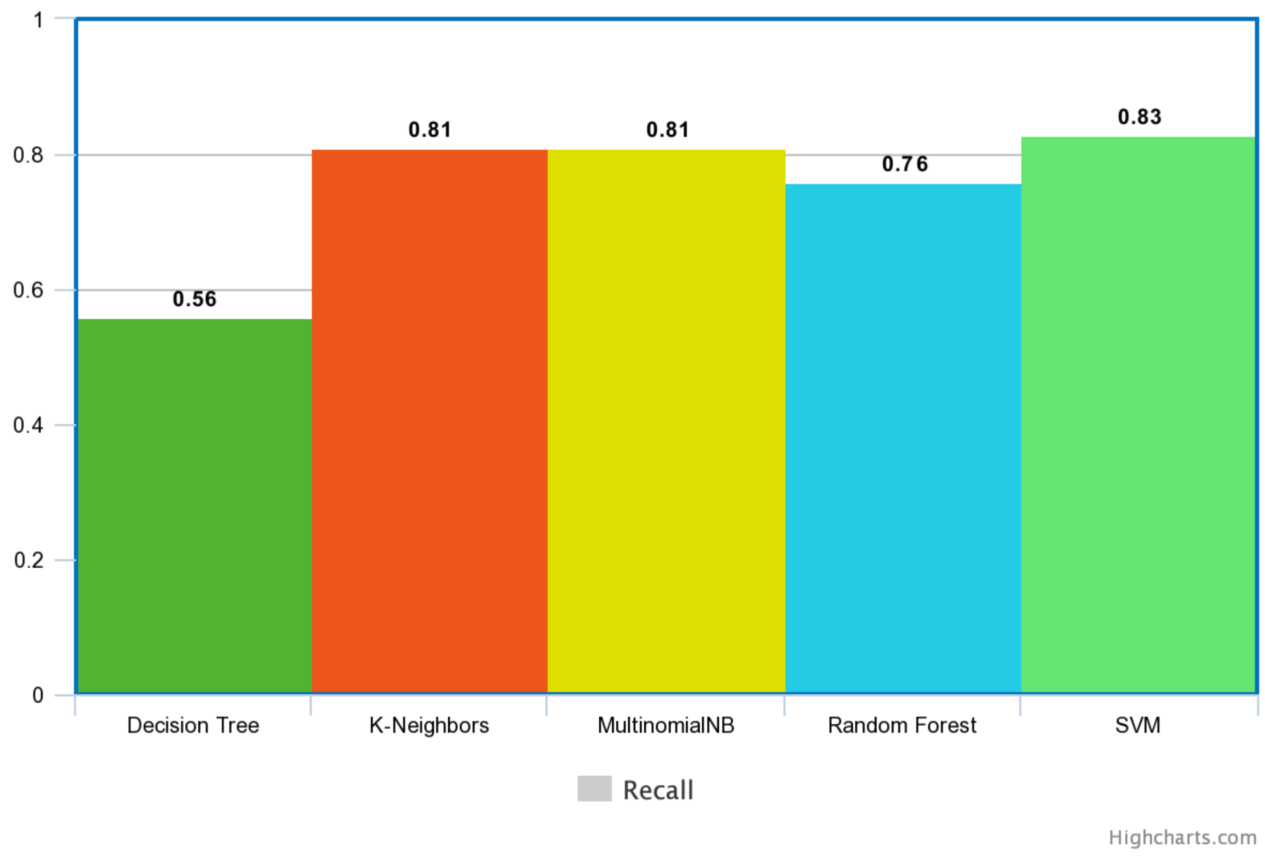
## Conclusion

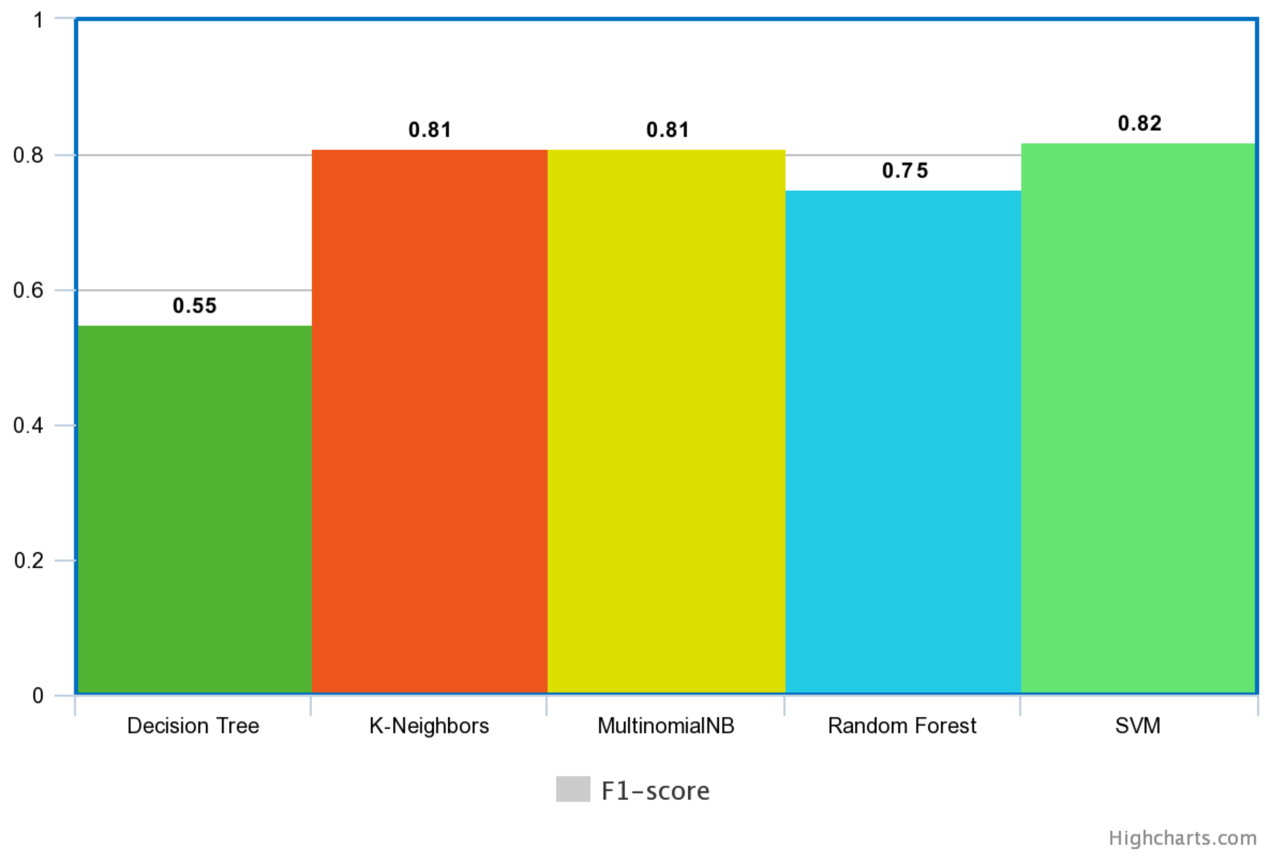


1
0.8
0.6
0.4
0.2
0
0.5555
0.81481
0.81488
0.76296
0.82593
Decision Tree
K-Neighbors
MultinomialNB
Random Forest
SVM
Accuracy
Highcharts.com



Untitled Chart
1
0.8
0.6
0.4
0.2
0
0.57
0.82
0.83
0.79
0.84
Decision Tree
K-Neighbors
MultinomialNB
Random Forest
SVM
Precision
Highcharts.com





As we studied at lesson, the best classifiers that show best performances on text mining are the SVM classifier (the best one) and the MultinomialNB classifier. Also the KNeighbors classifier, with the right choose of the k parameter, achieves great performances, similar to the SVM classifier. As regards the algorithm based on binary trees, we have bad performance from the Decision Tree classifier (55% of accuracy is a result not acceptable) and acceptable performances from the Random Forest classifier.

In conclusion we can say that classifiers as MultinomialNB, KNeighbors and SVM have satisfying performances (about 17/18% labelling error) and they could be used to classify articles with an acceptable precision. However, it must be said that this model is static, and the vocabulary of the newspaper articles will change over time, and consequently it is possible that the accuracy of the classifiers will reduce. To solve this problem it would be necessary to make the model dynamic, implementing some concept of Data Stream.