🏠 HOME

a HOPEFULLY-SMART NEWS AGGREGATOR

IN SUPREMÆ DIGNITATIS

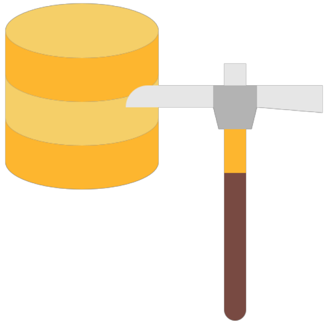
1343

Data Mining Project

Filippo Scotto

INTRODUCTION

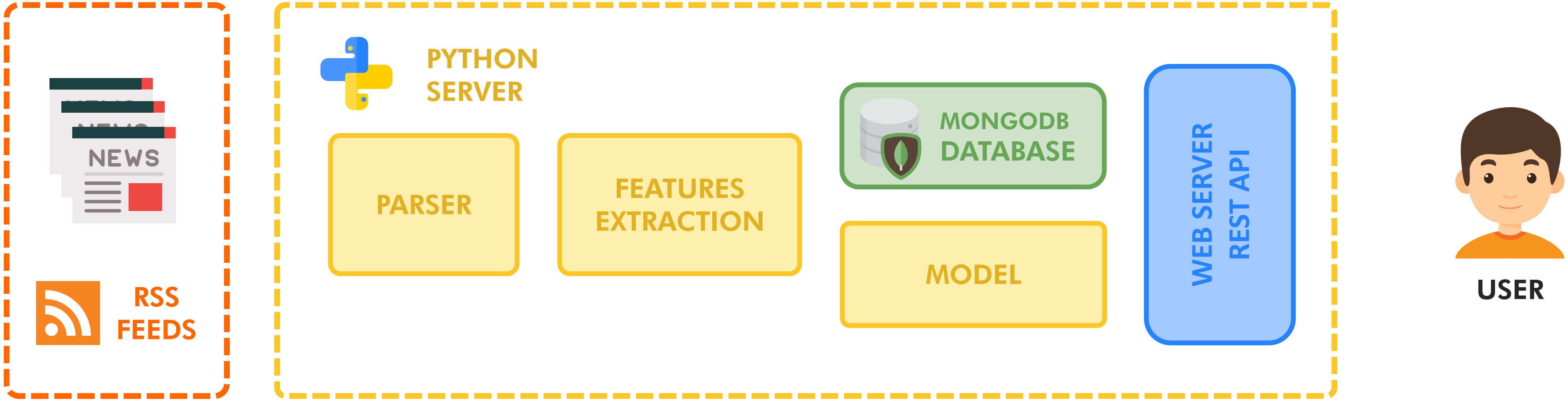
There are thousands of news sites, but most of their articles are *not for everybody*. Can we build a *machine learning based* system capable to **filter out** what we are not interesting in?



We need a system capable to classify the **category** of the articles and correctly predict their **likability**.

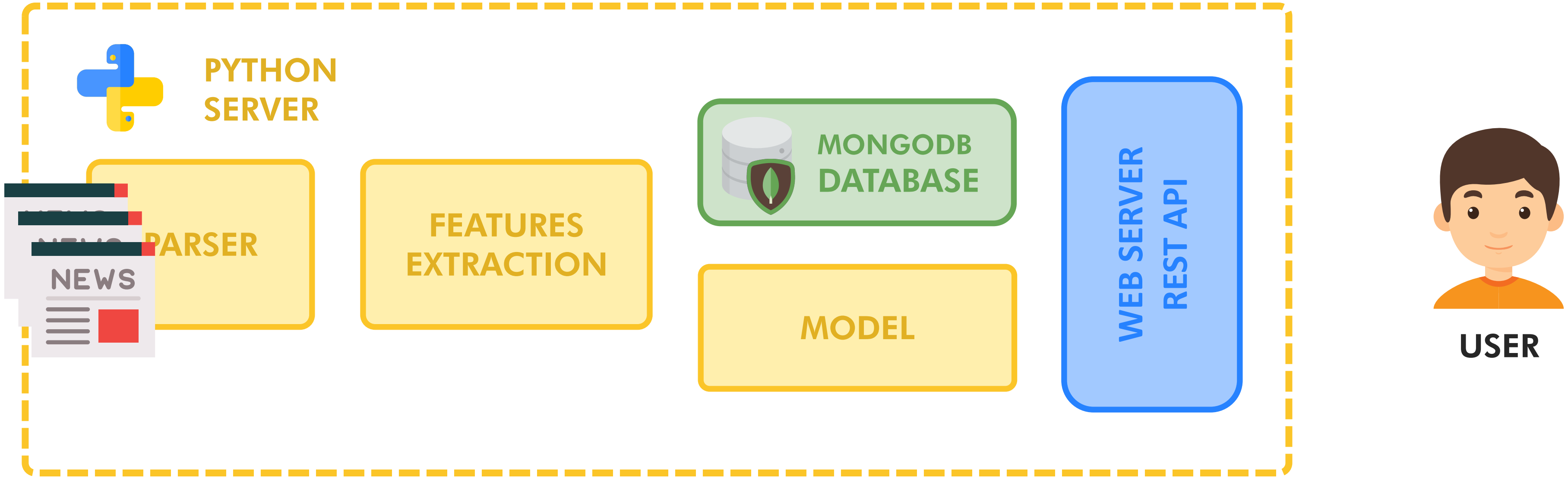
- The designed system will parse the **RSS feeds** coming from the most popular italian news sites;
- It will **preprocess** the data and **extract some features;**
- It will **classify** the news and predict wheter the user may like it or not.

THE SYSTEM



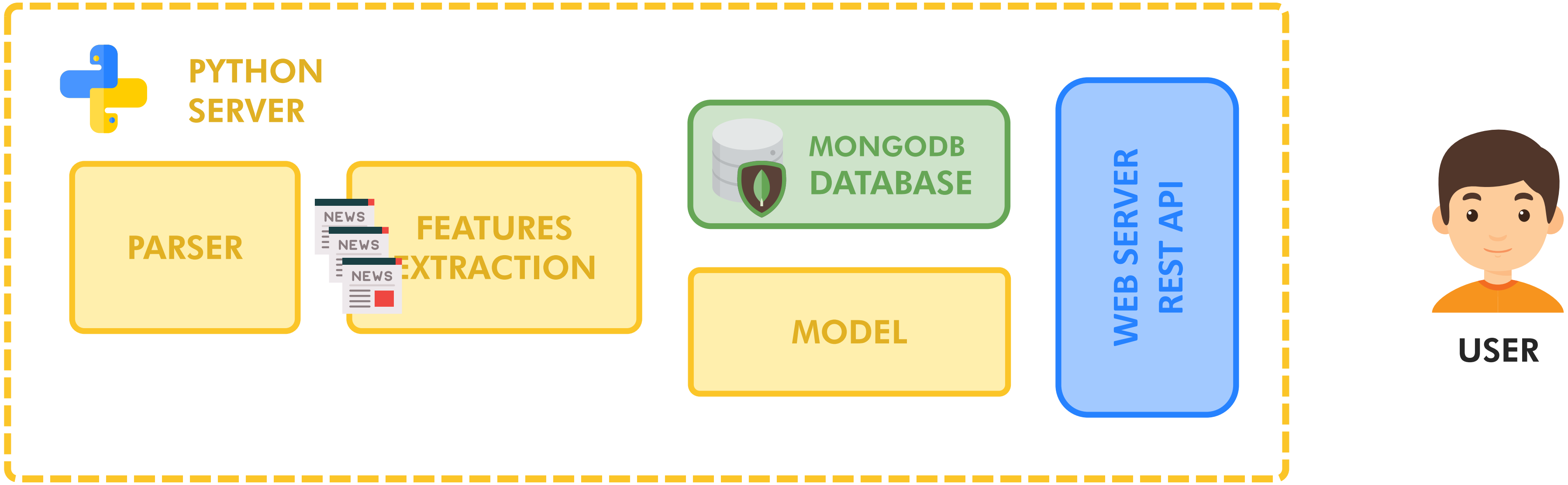
1 The articles are downloaded from the RSS Feeds

THE SYSTEM



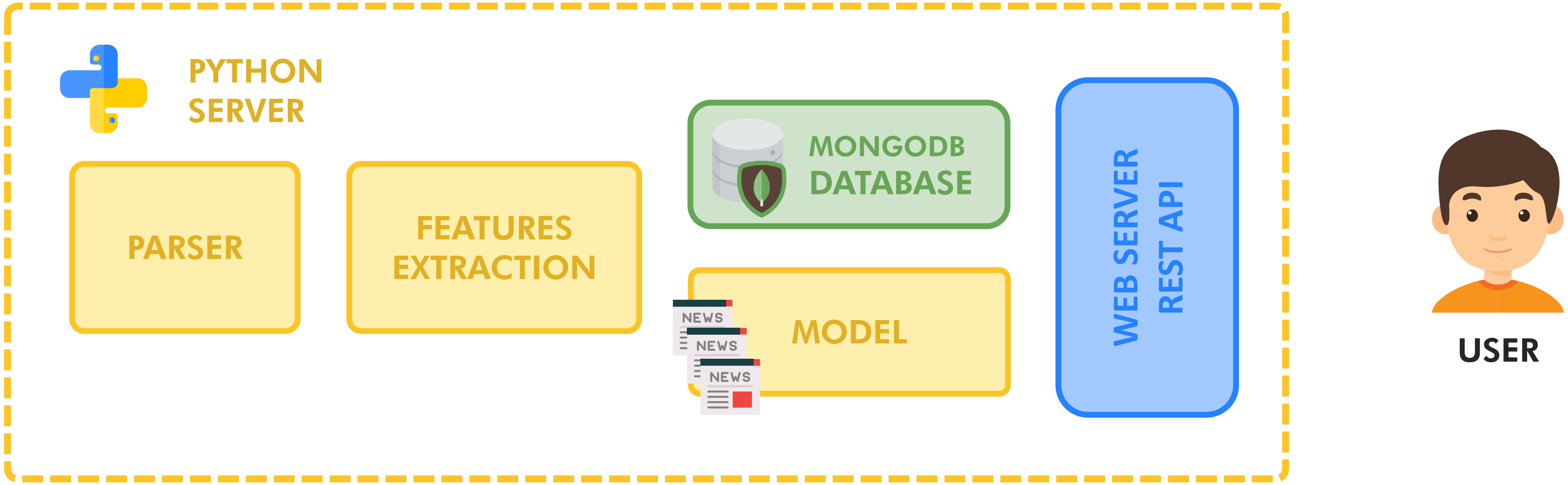
2 The articles are parsed by a Python Script

THE SYSTEM



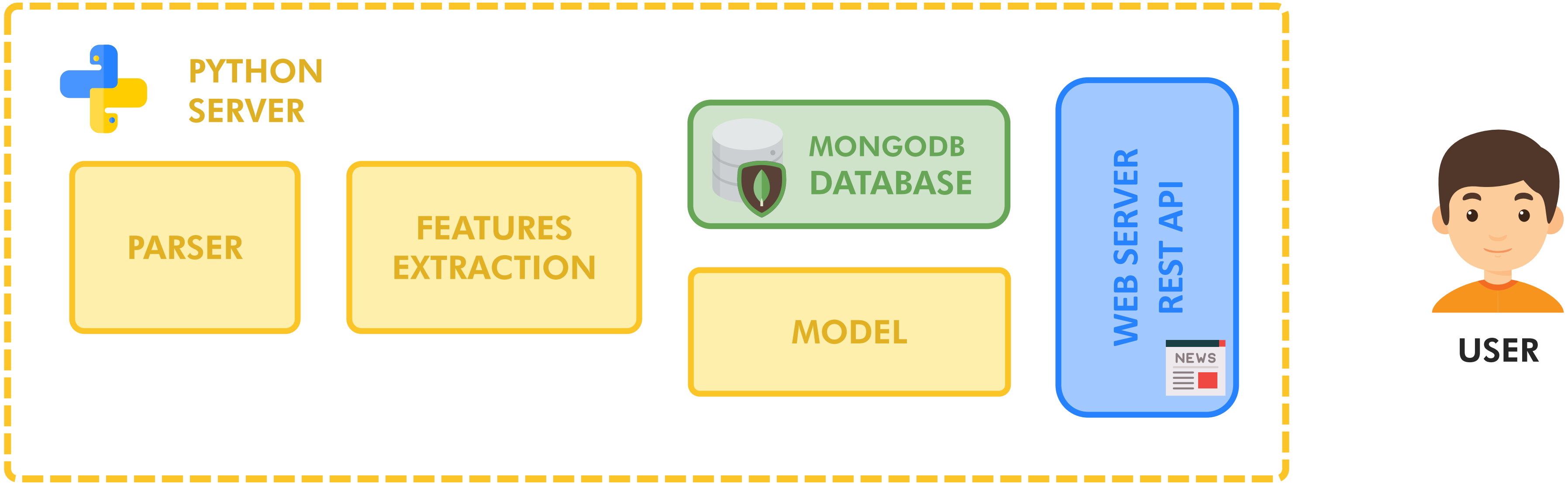
3 The parsed articles are processed to extract some featuers

THE SYSTEM



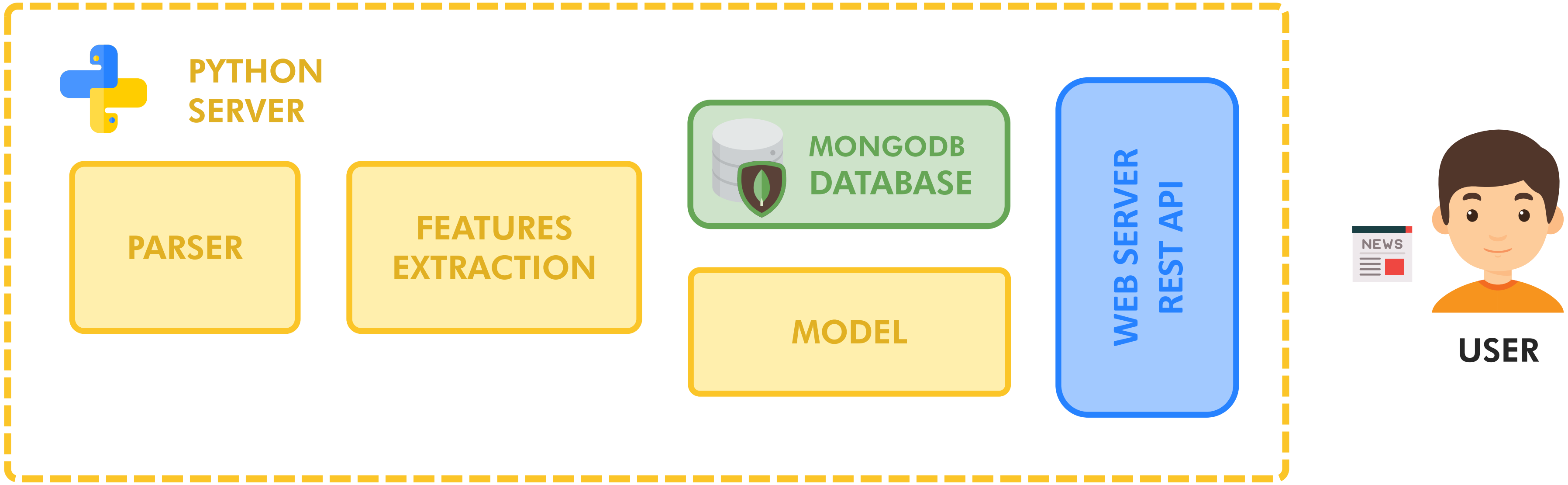
4 The features are given as input to a Machine Learning Model

THE SYSTEM



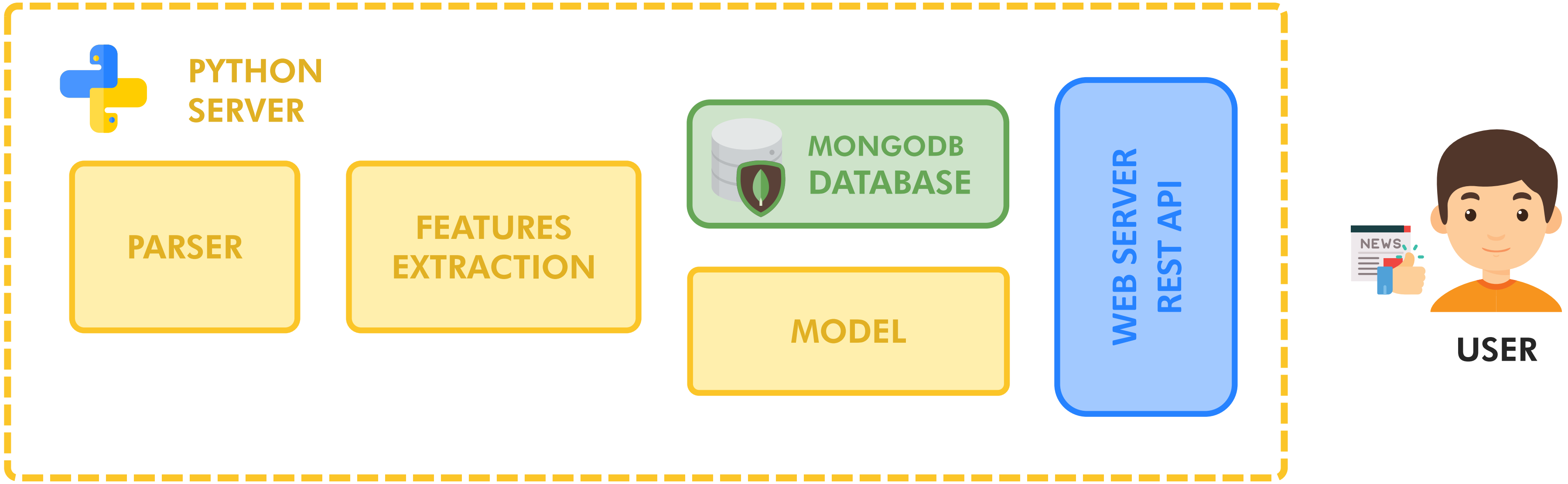
5 The filtered news are passed to the webserver

THE SYSTEM



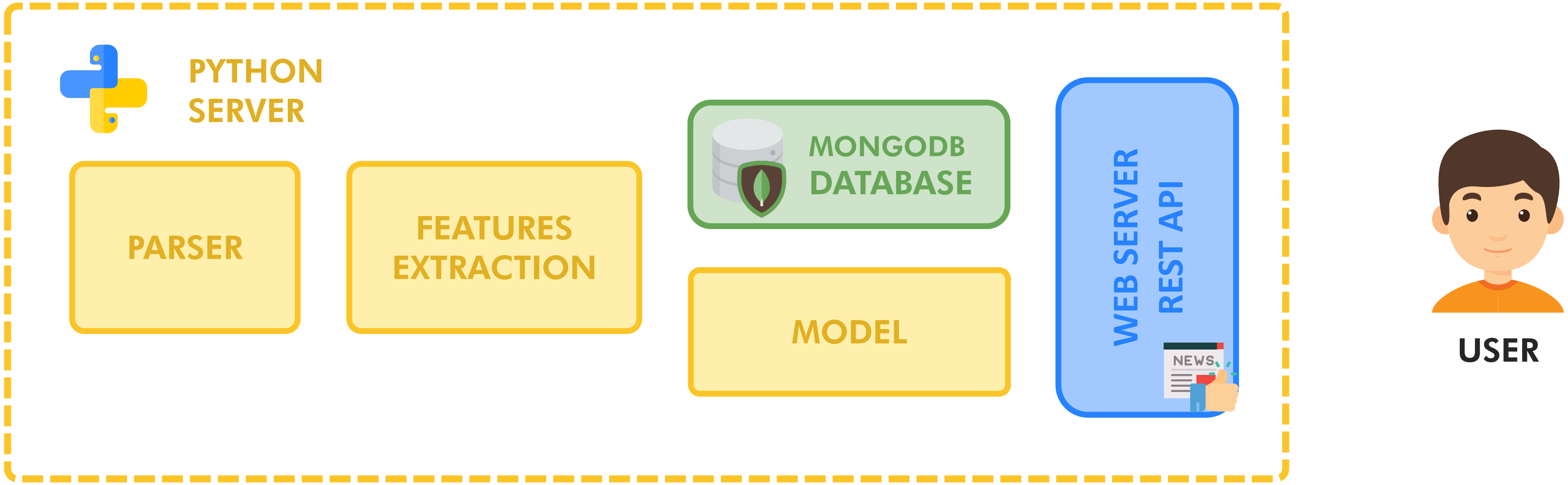
6 **The user can request its personalized feed using a REST API**

THE SYSTEM



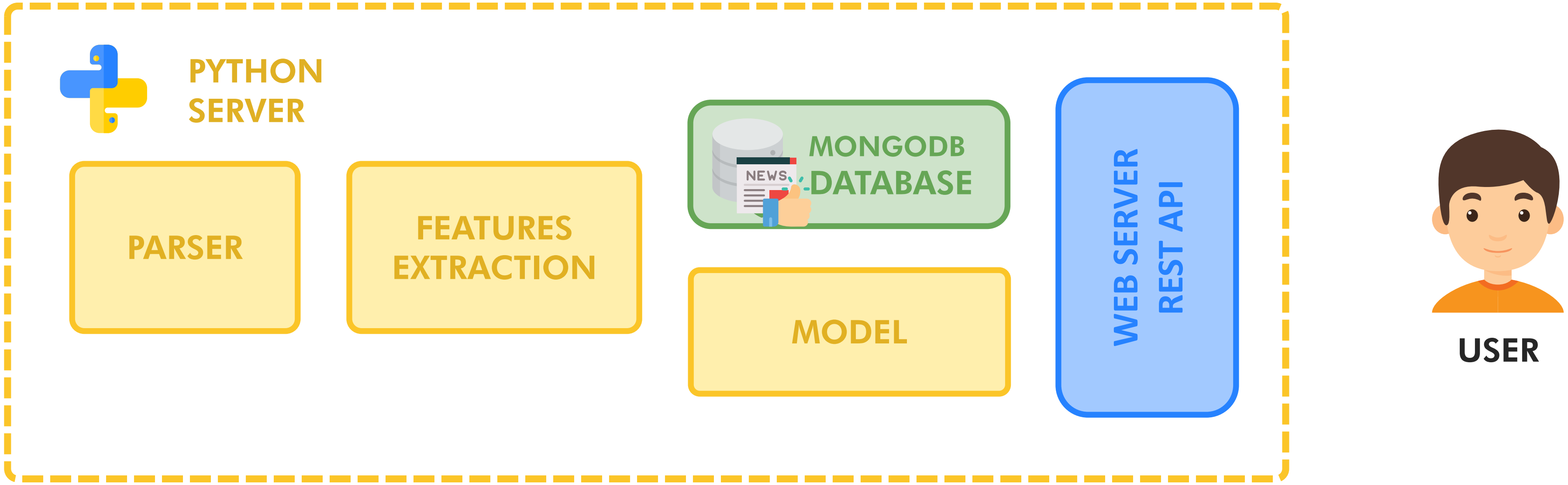
7 The user can “like” the article using the WebApp

THE SYSTEM



8 **The like request is processed by the WebServer**

THE SYSTEM

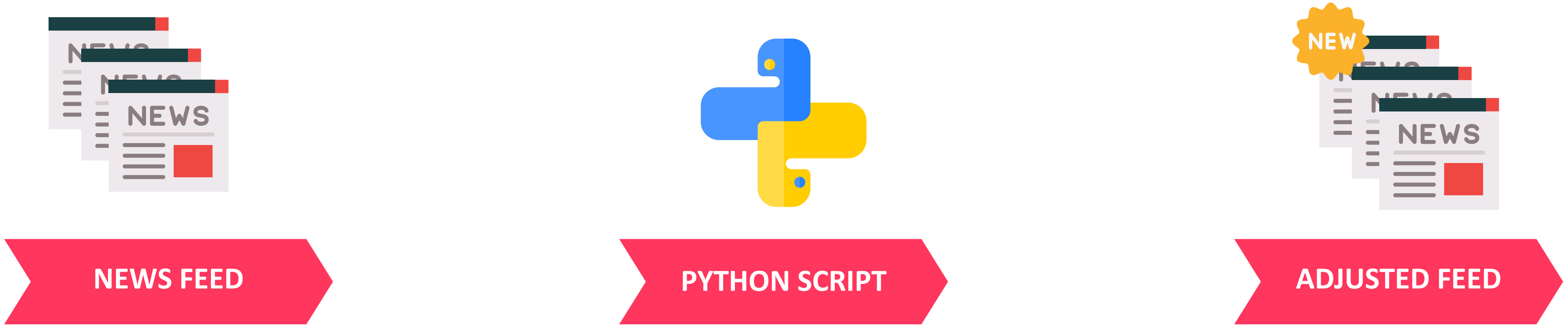


9 **Finally the database entry for the liked article is updated**

THE PARSER MODULE

News are collected from the RSS feeds from the most popular italian news sites. Unfortunately RSS feeds are not as structured as they were ment to be…

- **Missing values** and **different tag names**
- Article descriptions containing **images** or **raw HTML**
- We don’t need **everything**!



NEWS FEED

PYTHON SCRIPT

ADJUSTED FEED

THE PARSER MODULE: AN EXAMPLE

In the end what we obtain as output from the parser is a list of articles with the following structure:

Title
The title of the article

Author
The author of the article

Description
A *text-only snippet* of the content

Link
Hyperlink to the source

Timestamp
Published date-time

Image
Header image of the article

Missing *and* ***not valid values*** *will be replaced by an empty string!*

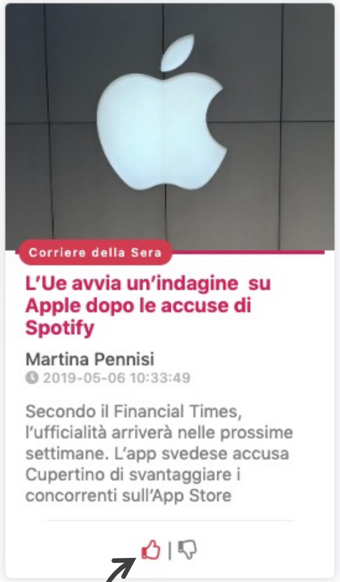
BUILDING THE DATASET

Once the feeds are parsed, we have to build a dataset for our model. In particular what we need to do is:

1. **Manually tag the articles**: each will be labelled with a category (nine different categories available)
2. **Manually Like/Dislike the articles**

To do this a **WebApp** was built: the app will show all the articles coming from the feed and then:

1. The article must be liked/disliked



Corriere della Sera

L'Ue avvia un'indagine su Apple dopo le accuse di Spotify

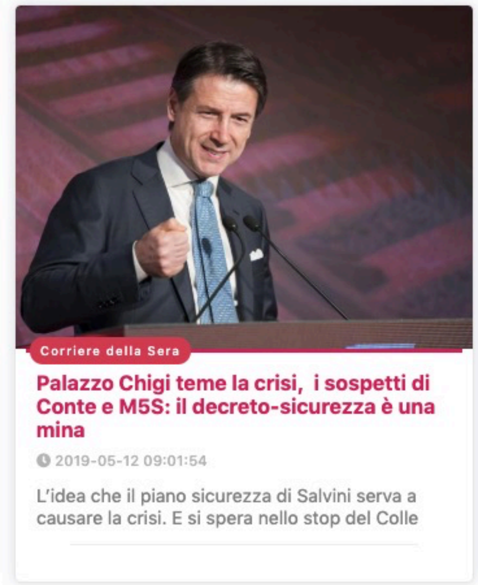
Martina Pennisi

2019-05-06 10:33:49

Secondo il Financial Times, l'ufficialità arriverà nelle prossime settimane. L'app svedese accusa Cupertino di svantaggiare i concorrenti sull'App Store

Like/Dislike buttons

2. The liked/disliked article will be inserted in the database and can be tagged



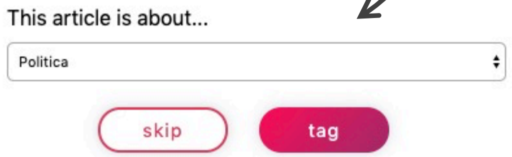
Corriere della Sera

Palazzo Chigi teme la crisi, i sospetti di Conte e M5S: il decreto-sicurezza è una mina

2019-05-12 09:01:54

L'idea che il piano sicurezza di Salvini serva a causare la crisi. E si spera nello stop del Colle

Form to tag an article



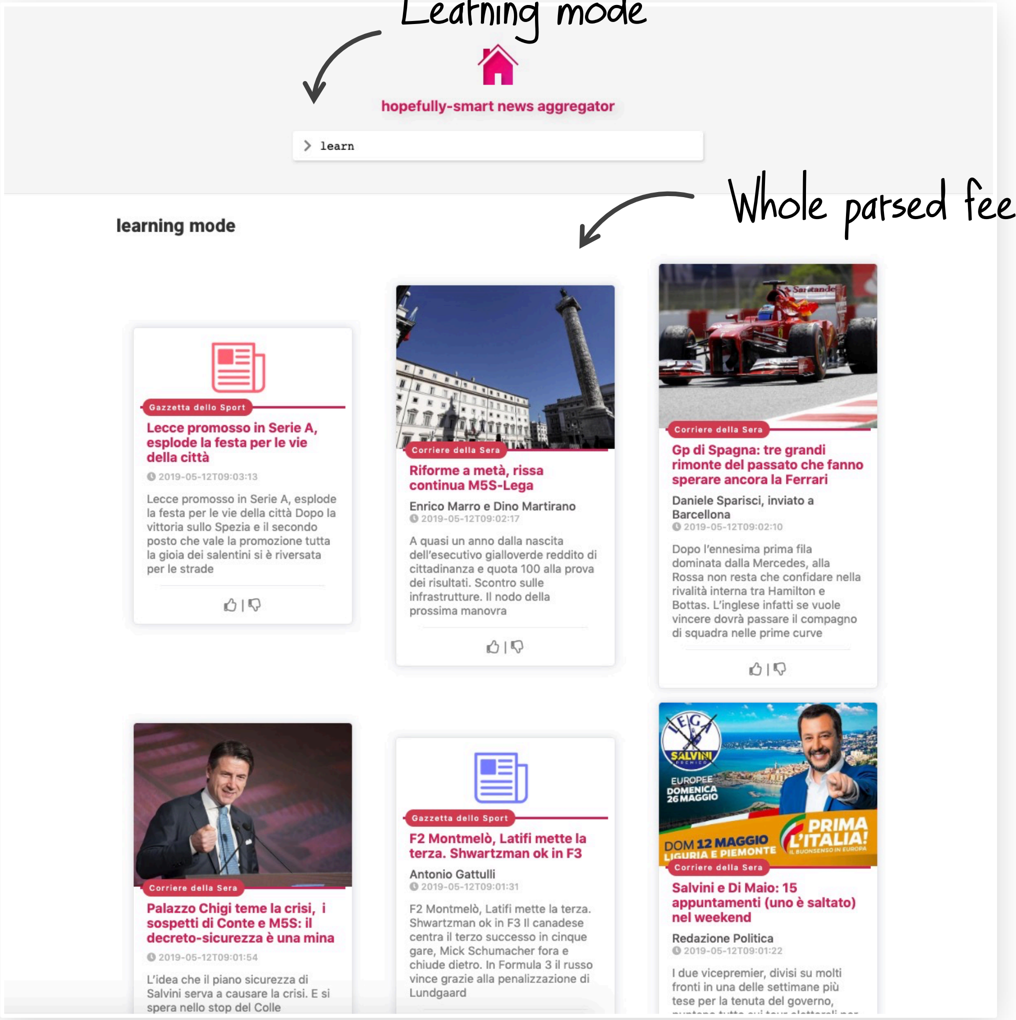
This article is about...

Politica

Button to skip current article and tag the next one

BUILDING THE DATASET: THE WEBAPP

Learning mode



hopefully-smart news aggregator

```
> learn
```

learning mode

Gazzetta dello Sport

Lecce promosso in Serie A, esplode la festa per le vie della città

2019-05-12T09:03:13

Lecce promosso in Serie A, esplode la festa per le vie della città Dopo la vittoria sullo Spezia e il secondo posto che vale la promozione tutta la gioia dei salentini si è riversata per le strade

👍 | 👎

Corriere della Sera

Riforme a metà, rissa continua M5S-Lega

Enrico Marro e Dino Martirano

2019-05-12T09:02:17

A quasi un anno dalla nascita dell'esecutivo gialloverde reddito di cittadinanza e quota 100 alla prova dei risultati. Scontro sulle infrastrutture. Il nodo della prossima manovra

👍 | 👎

Corriere della Sera

Gp di Spagna: tre grandi rimonte del passato che fanno sperare ancora la Ferrari

Daniele Sparisci, inviato a Barcellona

2019-05-12T09:02:10

Dopo l'ennesima prima fila dominata dalla Mercedes, alla Rossa non resta che confidare nella rivalità interna tra Hamilton e Bottas. L'inglese infatti se vuole vincere dovrà passare il compagno di squadra nelle prime curve

👍 | 👎

Corriere della Sera

Palazzo Chigi teme la crisi, i sospetti di Conte e M5S: il decreto-sicurezza è una mina

2019-05-12T09:01:54

L'idea che il piano sicurezza di Salvini serva a causare la crisi. E si spera nello stop del Colle

Gazzetta dello Sport

F2 Montmelò, Latifi mette la terza. Shwartzman ok in F3

Antonio Gattulli

2019-05-12T09:01:31

F2 Montmelò, Latifi mette la terza. Shwartzman ok in F3 Il canadese centra il terzo successo in cinque gare, Mick Schumacher fora e chiude dietro. In Formula 3 il russo vince grazie alla penalizzazione di Lundgaard

Corriere della Sera

Salvini e Di Maio: 15 appuntamenti (uno è saltato) nel weekend

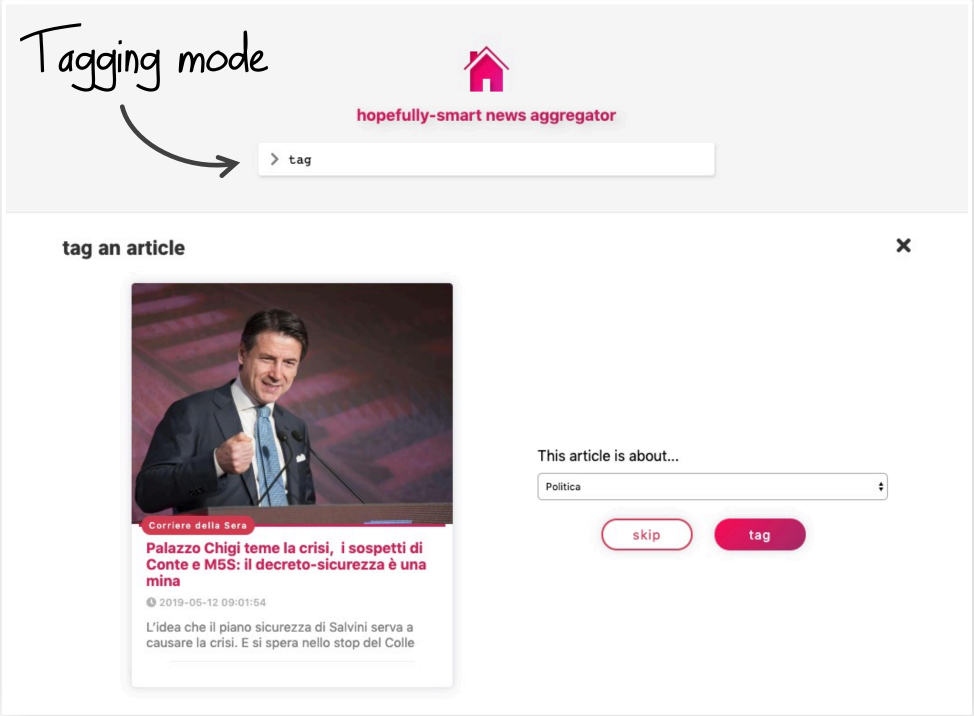
Redazione Politica

2019-05-12T09:01:22

I due vicepremier, divisi su molti fronti in una delle settimane più tese per la tenuta del governo,

Whole parsed feed

Tagging mode



hopefully-smart news aggregator

```
> tag
```

tag an article ✕

Corriere della Sera

Palazzo Chigi teme la crisi, i sospetti di Conte e M5S: il decreto-sicurezza è una mina

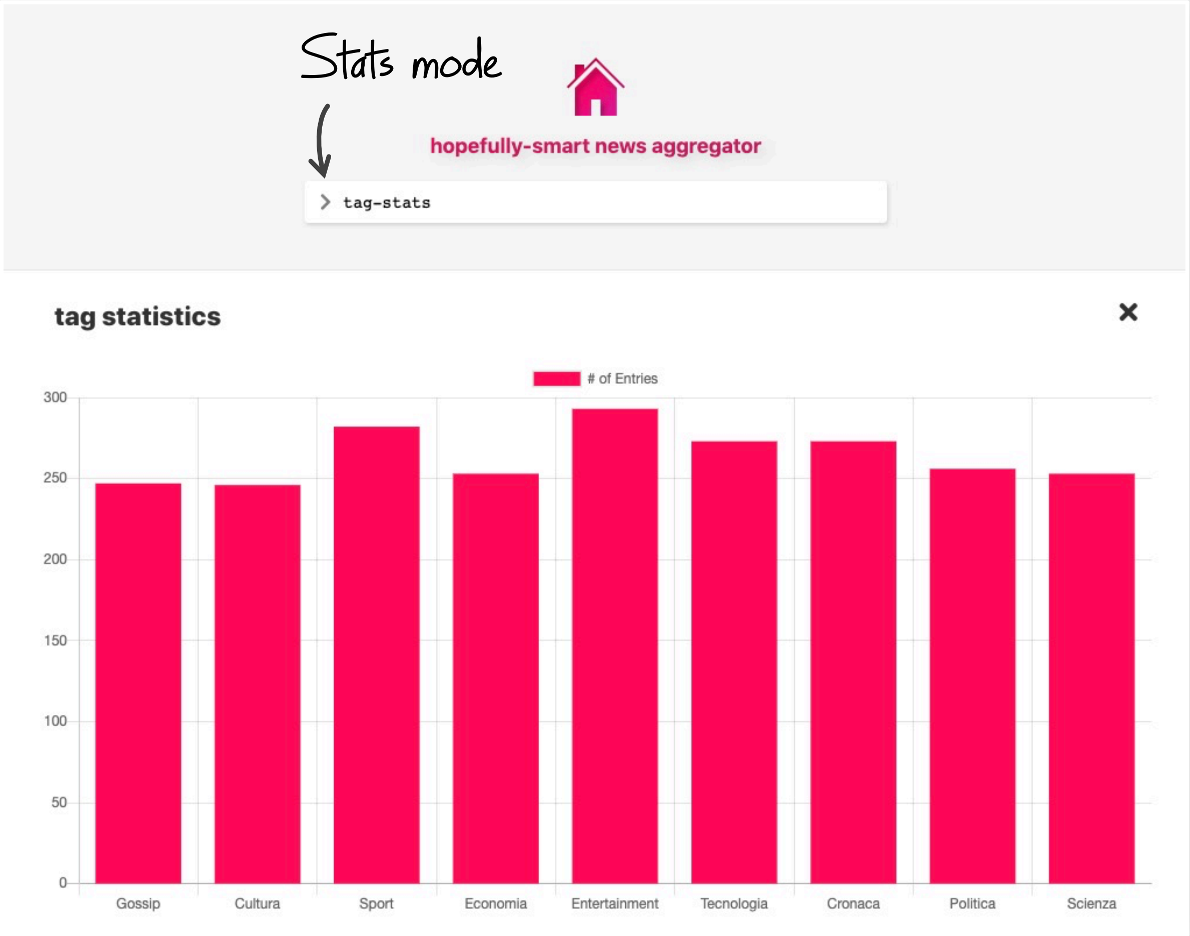
2019-05-12 09:01:54

L'idea che il piano sicurezza di Salvini serva a causare la crisi. E si spera nello stop del Colle

This article is about...

Politica

BUILDING THE DATASET: THE NEWS CATEGORIES



Stats mode

hopefully-smart news aggregator

```
> tag-stats
```

tag statistics

Tag distribution

Economy

Sport

Culture

Science

News

Entertainment

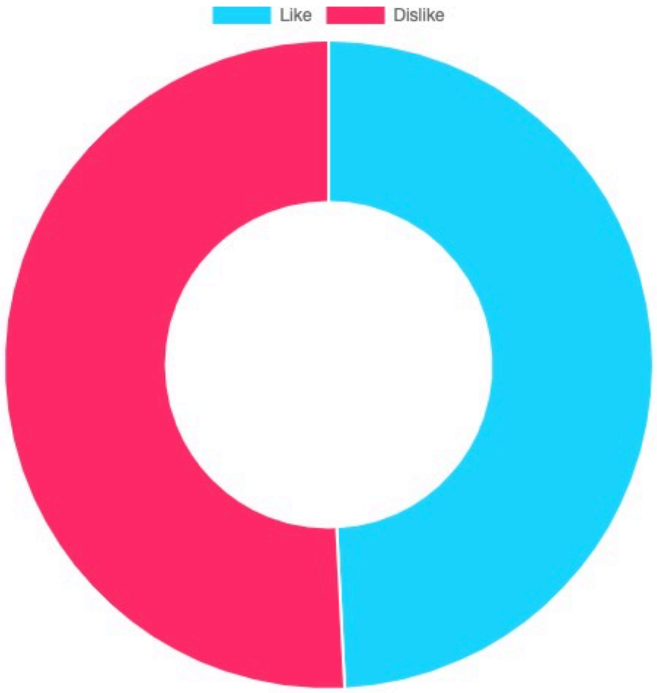
Gossip

Technology

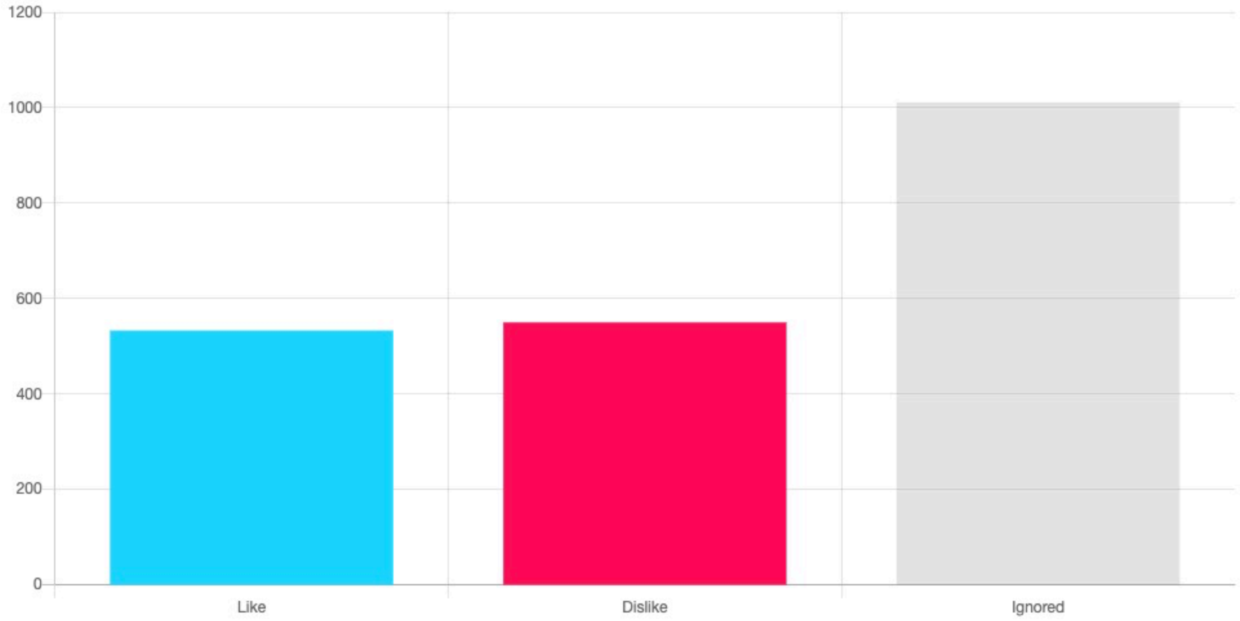
Politics

- **9 different categories**
- **3 months** news (since **March 2019**)
- Over **30 sources**
- **~ 250** articles per category

BUILDING THE DATASET: THE LIKE/DISLIKE DISTRIBUTION

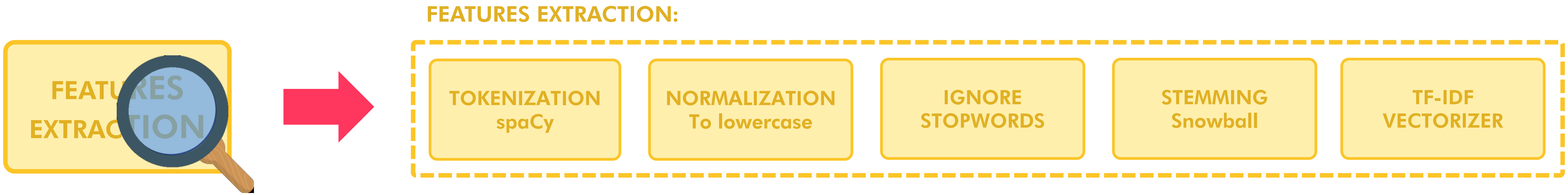


Like/Dislike distribution



Extra articles, useful to improve the category classification

EXTRACTING THE FEATURES



FEATURES EXTRACTION

FEATURES EXTRACTION:

TOKENIZATION spaCy

NORMALIZATION To lowercase

IGNORE STOPWORDS

STEMMING Snowball

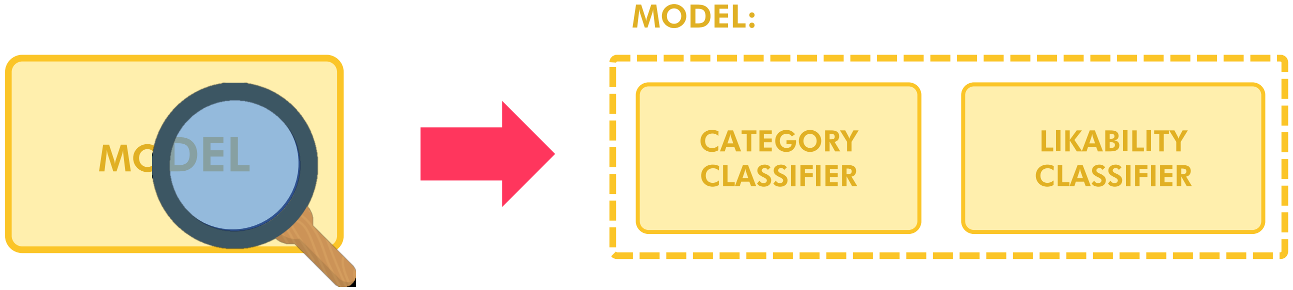
TF-IDF VECTORIZER

Now we need to mine the news in order to extract some useful features.

1. **Information Retrieval:**
    a. **Normalization**
    b. **Tokenization**
    c. **Stemming**
    d. **Ignoring Stopwords**
2. **Vectorizer**
    a. Build a vocabulary (**14500** top terms)
    b. Score normalization (**L2**)

Input: Article title + article description
Output: Sparse matrix containing the TF-IDF scores for the article's terms

BUILDING THE MODEL



The machine learning model that we are going to use it's actually composed by **two classifiers**:

CATEGORY CLASSIFIER

Predict the category for the article knowing the matrix of TF-IDF scores.

LIKABILITY CLASSIFIER

Predict wheter the user will like the article or not knowing the TF-IDF matrix and the category.

THE CATEGORY CLASSIFIER: BASE ESTIMATORS

Here are reported the **scores** and **performance** of some **simple classifiers** trying to predict the **category** of an article:

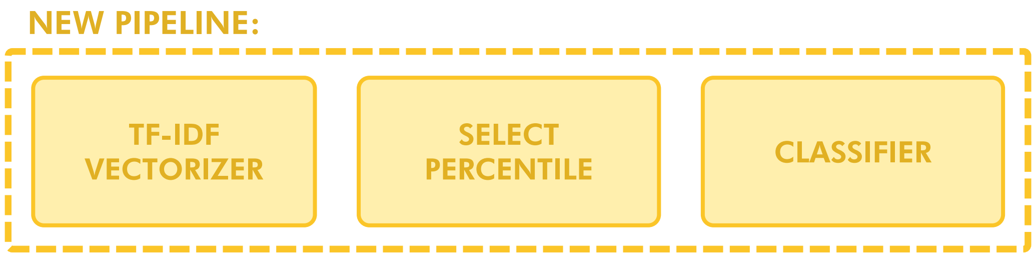
CLASSIFIER	F1-SCORE		ACCURACY		PRECISION		AUC ROC	
	MEAN	STD DEV	MEAN	STD DEV	MEAN	STD DEV	MEAN	STD DEV
C4.5 Decision Tree	0.5448	0.0294	0.5438	0.0283	0.5578	0.0326	0.7434	0.0158
MN-Bayes	0.7756	0.0344	0.7775	0.0346	0.7933	0.0340	0.8748	0.0194
LinearSVC	0.8031	0.0280	0.8032	0.0283	0.8105	0.0275	0.8893	0.0159
LogisticRegression	0.7914	0.0314	0.7918	0.0312	0.7995	0.0321	0.8828	0.0175

Best score!

The following classifiers were **tuned** using a GridSearch algorithm! The testing and validation were performed using **10 Folds Cross Validation**

BASE CLASSIFIER WITH FEATURES SELECTION (FILTER)

The pipeline model can be extended with a feature selectio method. Several attempts were done, best results were obtained using filter: **40% percentile** best features based on statistical tests (**chi2**)



NEW PIPELINE:

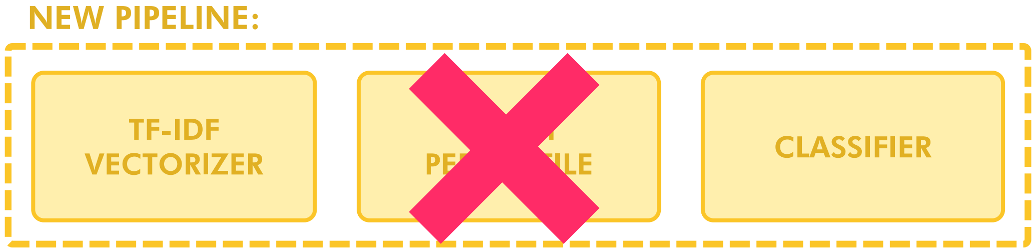
TF-IDF VECTORIZER

SELECT PERCENTILE

CLASSIFIER

BASE CLASSIFIER WITH FEATURES SELECTION

The pipeline model can be extended with a feature selectio method. Several attempts were done, best results were obtained using filter: **40% percentile** best features based on statistical tests (**chi2**)



NEW PIPELINE:

TF-IDF VECTORIZER

PE ILE

CLASSIFIER

However, as you can see results were not so good. So it was **removed** from the pipeline!

CLASSIFIER	F1-SCORE		ACCURACY		PRECISION		AUC ROC	
	MEAN	STD DEV	MEAN	STD DEV	MEAN	STD DEV	MEAN	STD DEV
C4.5 Decision Tree	0.5400	0.0288	0.5494	0.0275	0.5628	0.0313	0.7466	0.0153
MN-Bayes	0.7701	0.0380	0.7727	0.0378	0.7860	0.0388	0.8720	0.0213
LinearSVC	0.7870	0.0346	0.7877	0.0342	0.7926	0.0359	0.8805	0.0193
LogisticRegression	0.7825	0.0387	0.7833	0.0384	0.7907	0.0399	0.8779	0.0216

THE CATEGORY CLASSIFIER: ENSEMBLES

Here are reported the **scores** and **performance** of some **ensemble classifiers** trying to predict the **category** of an article:

CLASSIFIER	F1-SCORE		ACCURACY		PRECISION		AUC ROC	
	MEAN	STD DEV	MEAN	STD DEV	MEAN	STD DEV	MEAN	STD DEV
MNB + AdaBoost	0.4473	0.0291	0.4460	0.0267	0.7771	0.0452	0.6845	0.0150
Random Forest	0.6914	0.0376	0.6952	0.0357	0.6992	0.0377	0.8285	0.0200
Voting	0.7944	0.0319	0.7946	0.0318	0.8020	0.0315	0.8844	0.0179
SVC + Bagging	0.8023	0.0291	0.8024	0.0292	0.8096	0.0289	0.8888	0.0164

Best score

The following classifiers were **tuned** using a GridSearch algorithm! The testing and validation were performed using **10 Folds Cross Validation**

THE CATEGORY CLASSIFIER: PAIRED T-TEST

A single run is not enough, **five cross-validation runs** were performed for all of the previous classifiers and finally a **t-test** was performed on the **accuracy** score (using the *Weka Experimenter Tool*):

CLASSIFIER	C4.5	MNB	SVC	LogReg	AdaBoost	R. Forest	Voting	Bagging
C4.5 Decision Tree	0.53858	0.78027 V	0.79857 V	0.79162 V	0.44304 *	0.69245 V	0.79484 V	0.79865 V
MN-Bayes	0.53858 *	0.78027	0.79857 V	0.79162 V	0.44304 *	0.69245 *	0.79484 V	0.79865 V
LinearSVC	0.53858 *	0.78027 *	0.79857	0.79162 *	0.44304 *	0.69245 *	0.79484 *	0.79865
LogisticRegression	0.53858 *	0.78027 *	0.79857 V	0.79162	0.44304 *	0.69245 *	0.79484 V	0.79865 V
MNB + AdaBoost	0.53858 V	0.78027 V	0.79857 V	0.79162 V	0.44304	0.69245 V	0.79484 V	0.79865 V
Random Forest	0.53858 *	0.78027 V	0.79857 V	0.79162 V	0.44304 *	0.69245	0.79484 V	0.79865 V
Voting	0.53858 *	0.78027 *	0.79857 V	0.79162 *	0.44304 *	0.69245 *	0.79484	0.79865 V
SVC + Bagging	0.53858 *	0.78027 *	0.79857	0.79162 *	0.44304 *	0.69245 *	0.79484 *	0.79865

Best Classifier!

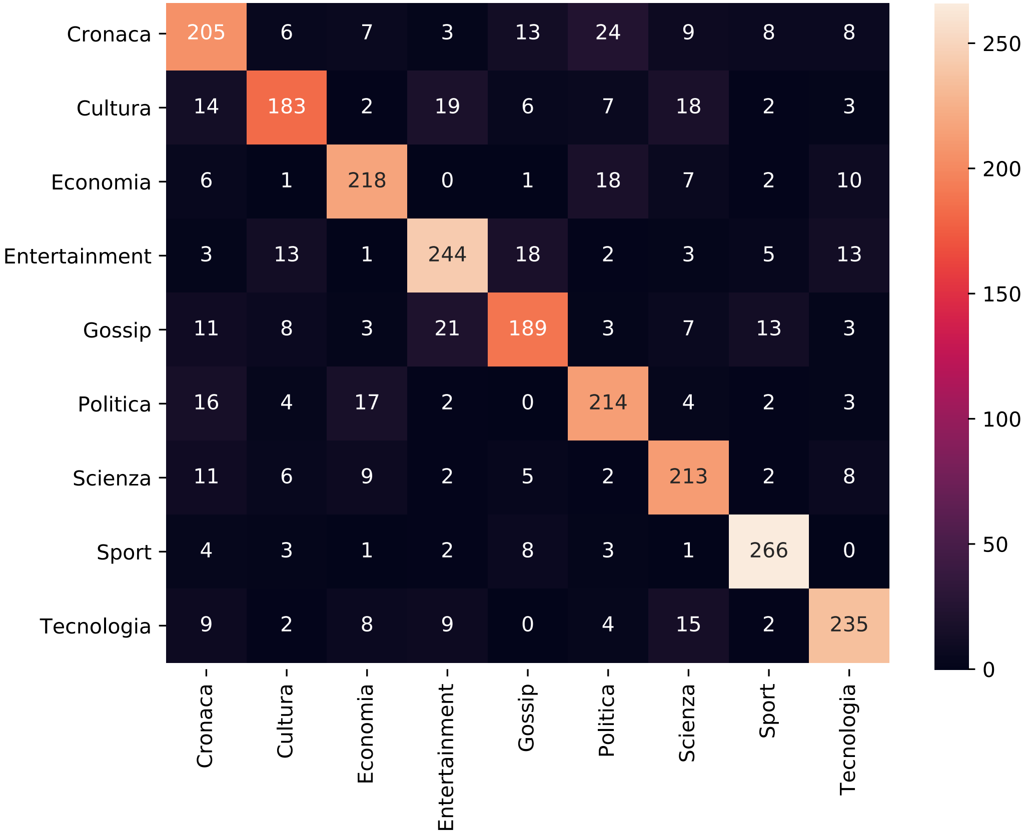
Confidence: 0.05

V : The results are statistically better

***** : The results are statistically worse

THE CATEGORY CLASSIFIER

In the end the best classifier was the **Bagging Classifier** so it was chosen as the classifier that should predict the article category. Here is shown its **confusion matrix** and some **parameters**:



	Cronaca	Cultura	Economia	Entertainment	Gossip	Politica	Scienza	Sport	Tecnologia
Cronaca	205	6	7	3	13	24	9	8	8
Cultura	14	183	2	19	6	7	18	2	3
Economia	6	1	218	0	1	18	7	2	10
Entertainment	3	13	1	244	18	2	3	5	13
Gossip	11	8	3	21	189	3	7	13	3
Politica	16	4	17	2	0	214	4	2	3
Scienza	11	6	9	2	5	2	213	2	8
Sport	4	3	1	2	8	3	1	266	0
Tecnologia	9	2	8	9	0	4	15	2	235

```
parameters {
  'base': LinearSVC(),
  'n_estimators': 100
}
```

The chosen classifier has an 80% accuracy, is this good?
Indeed it could be better, however predicting the category of an article is not an easy task (not even for a human), so in the end we can say that it is ***fair enough.***

THE LIKABILITY CLASSIFIER: BASE ESTIMATORS

Here are reported the **scores** and **performance** of some **simple classifiers** trying to predict the **likability** of an article:

CLASSIFIER	F1-SCORE		ACCURACY		PRECISION		AUC ROC	
	MEAN	STD DEV	MEAN	STD DEV	MEAN	STD DEV	MEAN	STD DEV
Decision Tree	0.7145	0.0366	0.7147	0.0365	0.7154	0.0365	0.7146	0.0365
MN-Bayes	0.8607	0.0244	0.8615	0.0241	0.8680	0.0242	0.8605	0.0242
LinearSVC	0.8715	0.0334	0.8716	0.0334	0.8720	0.0338	0.8717	0.0333
LogisticRegression	0.8725	0.0312	0.8726	0.0312	0.8739	0.0316	0.8726	0.0312

Best score!

The following classifiers were **tuned** using a GridSearch algorithm! The testing and validation were performed using **10 Folds Cross Validation**

THE LIKABILITY CLASSIFIER: ENSEMBLES

Here are reported the **scores** and **performance** of some **ensemble classifiers** trying to predict the **category** of an article:

CLASSIFIER	F1-SCORE		ACCURACY		PRECISION		AUC ROC	
	MEAN	STD DEV	MEAN	STD DEV	MEAN	STD DEV	MEAN	STD DEV
MNB + ADABoost	0.3246	0.0025	0.4921	0.0023	0.2422	0.0022	0.5000	0.0001
Random Forest	0.7976	0.0312	0.7996	0.0302	0.8087	0.0271	0.7982	0.0306
Voting	0.8696	0.0355	0.8698	0.0355	0.8717	0.0359	0.8694	0.0355
SVC + Bagging	0.8669	0.0286	0.8670	0.0286	0.8690	0.0287	0.8672	0.0285

Best score

The following classifiers were **tuned** using a GridSearch algorithm! The testing and validation were performed using **10 Folds Cross Validation**

THE LIKABILITY CLASSIFIER: PAIRED T-TEST

Again, a single run is not enough, **five cross-validation runs** were performed for all of the previous classifiers and finally a **t-test** was performed on the **accuracy** score (using the *Weka Experimenter Tool*):

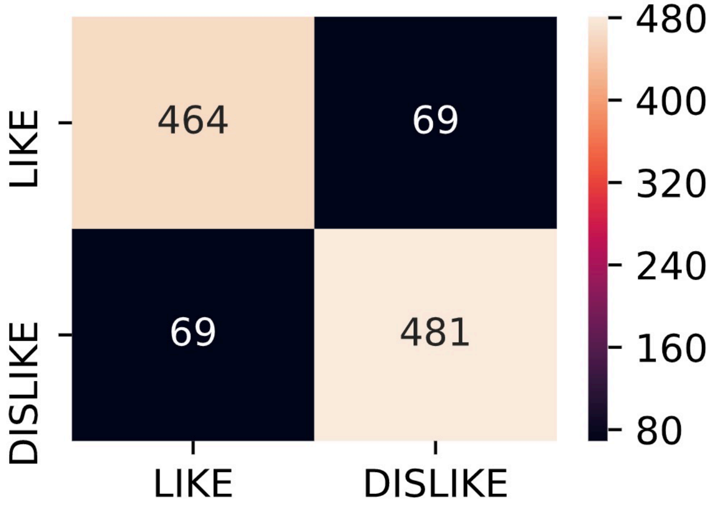
CLASSIFIER	C4.5	MNB	SVC	LogReg	AdaBoost	R. Forest	Voting	Bagging
C4.5 Decision Tree	0.71725	0.86503 V	0.86963 V	0.86853 V	0.49214 *	0.80462 V	0.86632 V	0.86724 V
MN-Bayes	0.71725 *	0.86503	0.86963	0.86853	0.49214 *	0.80462 *	0.86632	0.86724 V
LinearSVC	0.71725 *	0.86503	0.86963	0.86853	0.49214 *	0.80462 *	0.86632	0.86724
LogisticRegression	0.71725 *	0.86503	0.86963	0.86853	0.49214 *	0.80462 *	0.86632	0.86724
MNB + AdaBoost	0.71725 V	0.86503 V	0.86963 V	0.86853 V	0.49214	0.80462 V	0.86632 V	0.86724 V
Random Forest	0.71725 *	0.86503 V	0.86963 V	0.86853 V	0.49214 *	0.80462	0.86632 V	0.86724 V
Voting	0.71725 *	0.86503	0.86963	0.86853	0.49214 *	0.80462 *	0.86632	0.86724
SVC + Bagging	0.71725 *	0.86503	0.86963	0.86853	0.49214 *	0.80462 *	0.86632	0.86724

Confidence: 0.05
V : The results are statistically better
***** : The results are statistically worse

As you can see in this case there is no clear winner!
BaggingClassifier is still pretty strong,
and LogReg too!

THE LIKABILITY CLASSIFIER

LogisticRegression was chosen as the classifier that should predict the article likability. Here is shown its **confusion matrix**:



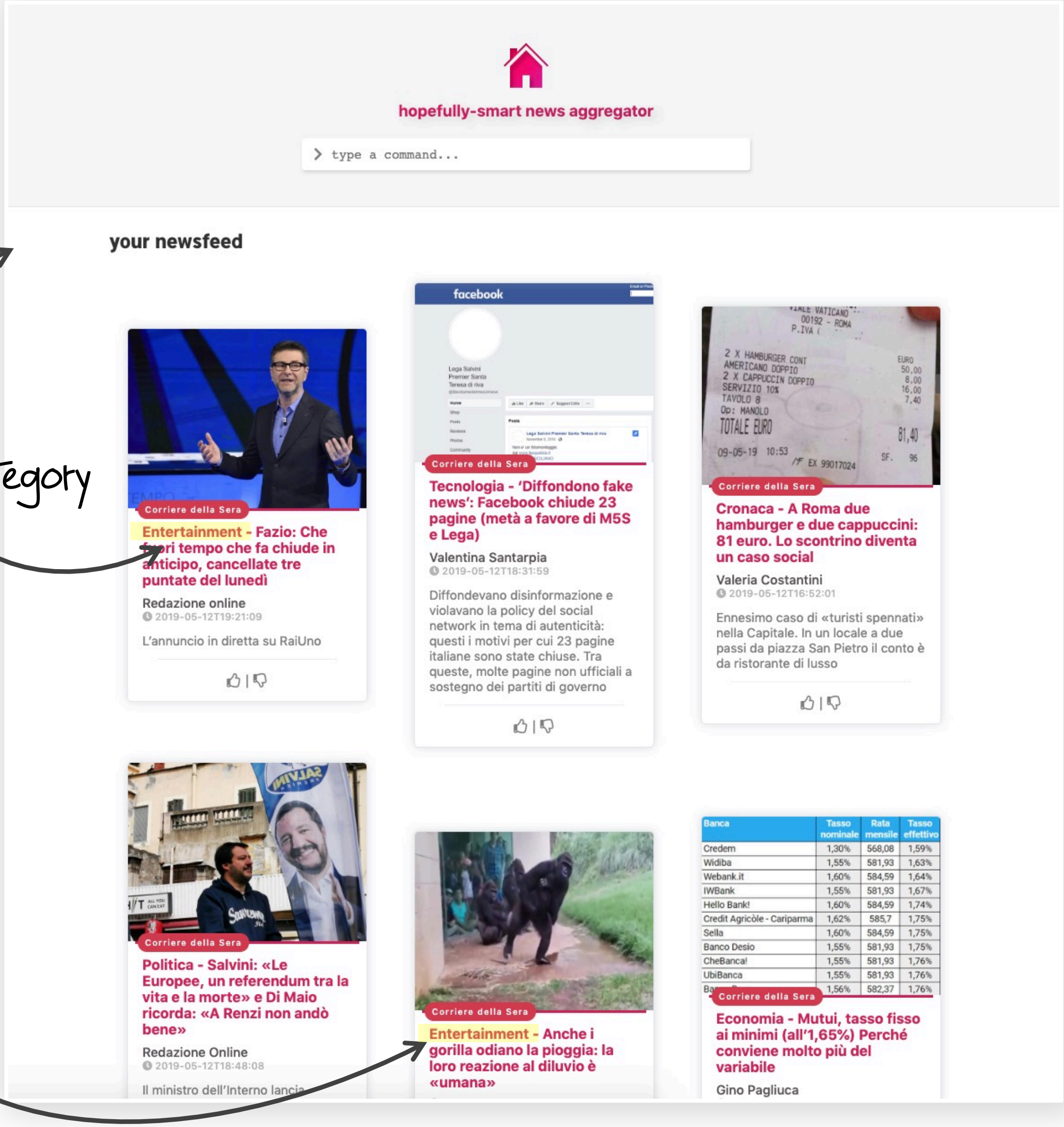
Please notice that this result was achieved under the assumption that the category of the article was correctly predicted !

THE WEB APPLICATION: A SCREENSHOT OF THE NEWSFEED

Personalized feed

Predicted category

Maybe an error?



your newsfeed

Corriere della Sera

Entertainment - Fazio: Che fuori tempo che fa chiude in anticipo, cancellate tre puntate del lunedì

Redazione online
2019-05-12T19:21:09

L'annuncio in diretta su RaiUno

Corriere della Sera

Tecnologia - 'Diffondono fake news': Facebook chiude 23 pagine (metà a favore di M5S e Lega)

Valentina Santarpia
2019-05-12T18:31:59

Diffondevano disinformazione e violavano la policy del social network in tema di autenticità: questi i motivi per cui 23 pagine italiane sono state chiuse. Tra queste, molte pagine non ufficiali a sostegno dei partiti di governo

Corriere della Sera

Cronaca - A Roma due hamburger e due cappuccini: 81 euro. Lo scontrino diventa un caso social

Valeria Costantini
2019-05-12T16:52:01

Ennesimo caso di «turisti spennati» nella Capitale. In un locale a due passi da piazza San Pietro il conto è da ristorante di lusso

Corriere della Sera

Politica - Salvini: «Le Europee, un referendum tra la vita e la morte» e Di Maio ricorda: «A Renzi non andò bene»

Redazione Online
2019-05-12T18:48:08

Il ministro dell'Interno lancia

Corriere della Sera

Entertainment - Anche i gorilla odiano la pioggia: la loro reazione al diluvio è «umana»

Banca	Tasso nominale	Rata mensile	Tasso effettivo
Credem	1,30%	568,08	1,59%
Widiba	1,55%	581,93	1,63%
Webank.it	1,60%	584,59	1,64%
IWBank	1,55%	581,93	1,67%
Hello Bank!	1,60%	584,59	1,74%
Credit Agricòle - Cariparma	1,62%	585,7	1,75%
Sella	1,60%	584,59	1,75%
Banco Desio	1,55%	581,93	1,75%
CheBanca!	1,55%	581,93	1,76%
UbiBanca	1,55%	581,93	1,76%
Ba	1,56%	582,37	1,76%

Corriere della Sera

Economia - Mutui, tasso fisso ai minimi (all'1,65%) Perché conviene molto più del variabile

Gino Pagliuca

THANK YOU

… any questions?