Task2&3 -EDA, Classification

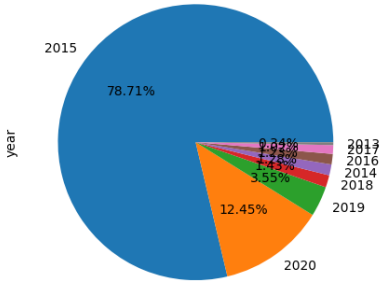
We start the process by: Using the Hespress Dataset NLP, particularly the stories, to perform exploratory data analysis, gain insights, and comprehend the data properly.

1. Text Preprocessing

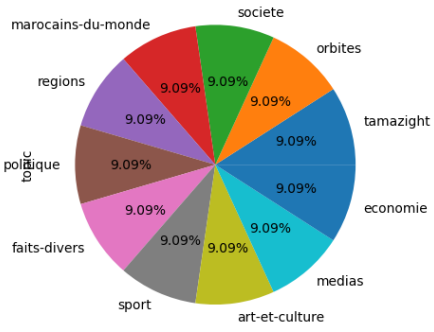
- Tokanization
- Remove Non-Arabic words.
- Remove Tashkeel
- Remove Punctuations
- Remove Digits
- Remove URLS if found in the text.

2. Exploratory Data Analysis

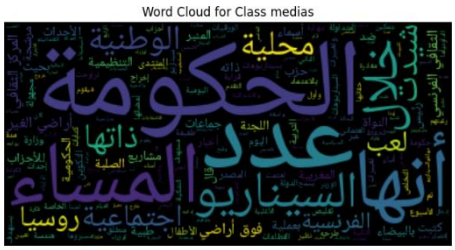
- **Year distribution**
  By using the date info in the dataset, we can get the year in which a lot of news was published and vice versa, This indicates that the events of a major event occurred in this year, and therefore there is a lot of text about it can be a benefit in classification by Focus on topics this year.
  According to the figure: 2015: It is the highest percentage of events.

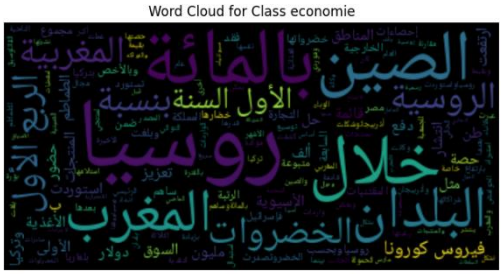


- **Number of examples per class**
  According to the number of topics in the dataset we can get the distribution of each topic or class and calculate the percentage of news in each class.
  The graph demonstrates that each class's distribution of classes or themes is equal. And the data is not biased to a specific topic.

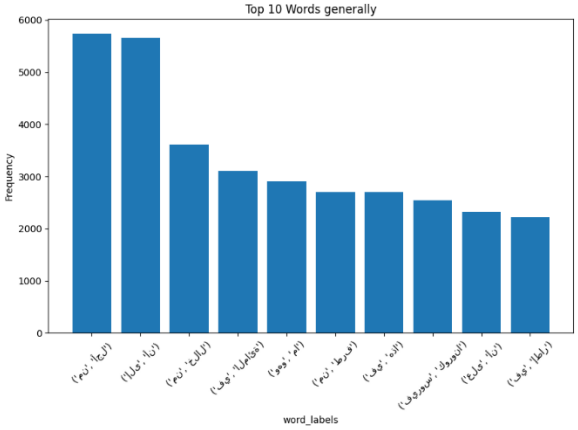


- **Wordcloud for top words per topic**
  From a plot that display a random and messy words from each topic, these words appeared sequentially and repeatedly and these words also express the type of each topic, for example in sports, by mentioning a club or match ,trainee,and words that refer to sport.

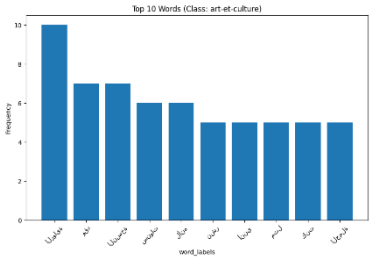


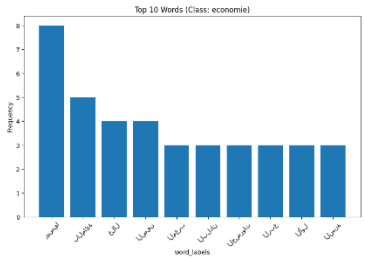


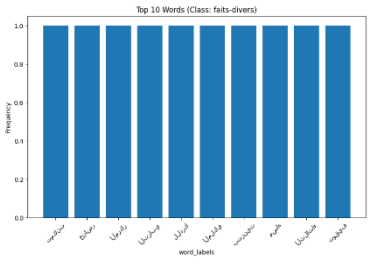
- **10 Top frequent 2 -gram generally**
  Here we get the top 10 words that are frequently mentioned together in all the dataset. We used 2 – gram to get the most 2 words sequentially appeared.



- **10 Top frequent 1-gram per class**
  In this part we get the top 10 words has the most count and frequent in each topic separately.

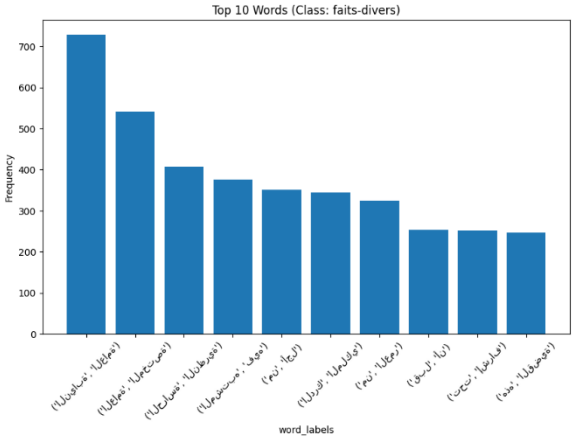


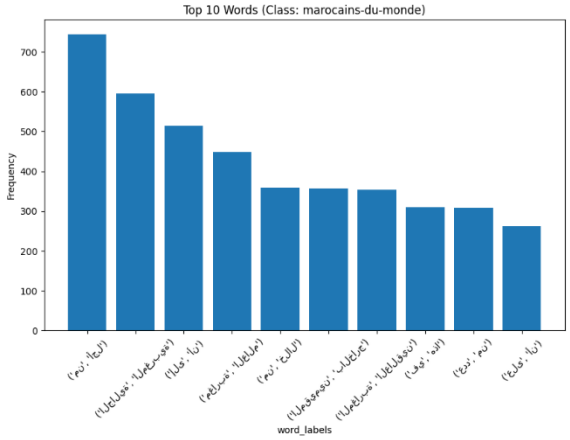




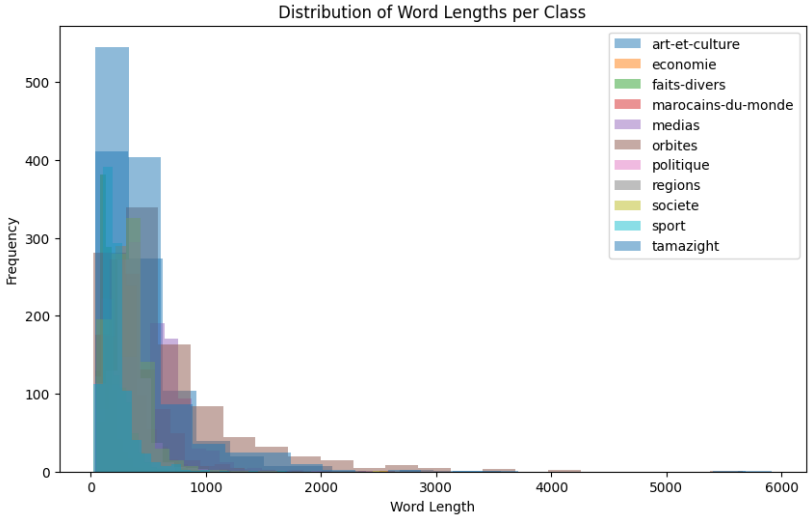
- **10 Top frequent 2 -gram per class**

We identified the top 10 double words per class that frequently appeared in the same context.





- **Distribution of Word Lengths per Class**
  To follow the news in depth we can calculate the distribution of sentence length per each topic and get the most topic that have more word and length.
  Form the figure: the average length ranges start from 0 to 1000

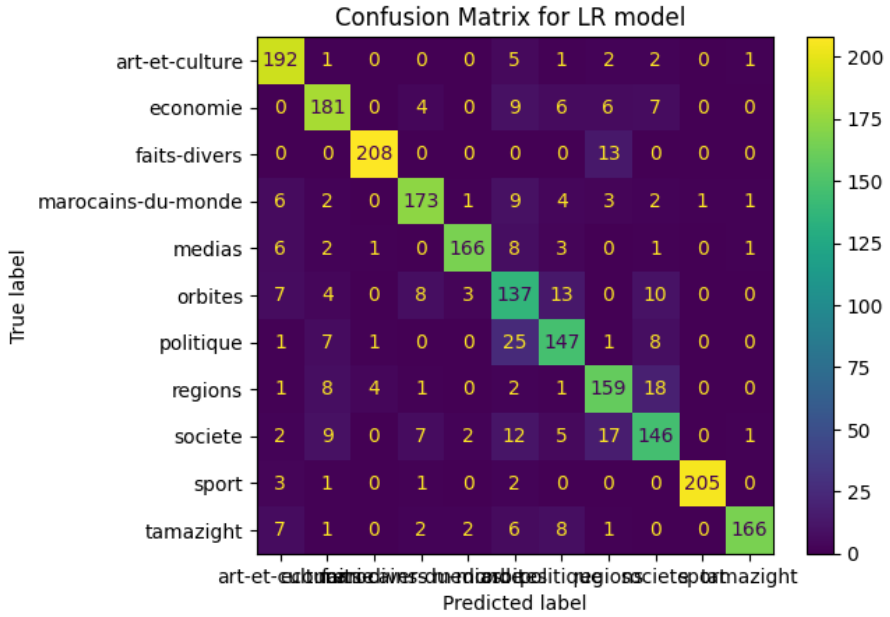


3. Feature Extraction

This dataset contains text so that we can manipulate this text, we must convert it into numerical values before modeling, we used TF_IDF to get the feature from the text and classify which text belongs to the topic.
TF_IDF deals in a good and direct way with text that can get the unique words and give these words more importance and it is best for Herpress data which in each topic contains different words expressing the topic

4. Modelling

after divide the dataset into 80% for training and 20% for testing sets. The training set is used to train the model, while the testing set is used to evaluate its performance on unseen data. At the stage we used classification model Logistic Regression to classify the topics, the accuracy of the LR is 85.45% indicating that it correctly classifies 85% of the instances in the test set, also by displaying confusion matric and classification report.



	precision	recall	f1-score	support
art-et-culture	0.85	0.94	0.90	204
economie	0.84	0.85	0.84	213
faits-divers	0.97	0.94	0.96	221
marocains-du-monde	0.88	0.86	0.87	202
medias	0.95	0.88	0.92	188
orbites	0.64	0.75	0.69	182
politique	0.78	0.77	0.78	190
regions	0.79	0.82	0.80	194
societe	0.75	0.73	0.74	201
sport	1.00	0.97	0.98	212
tamazight	0.98	0.86	0.91	193
accuracy			0.85	2200
macro avg	0.86	0.85	0.85	2200
weighted avg	0.86	0.85	0.86	2200

5. Model Evaluation

The provided figure appears the classification report for the classifier of multi class in display accuracy, recall, precision, and f1 score for each class. These evaluation matrices indicate the model goes well way in classification.

In precision matrix that calculates the predicted correctly among all predicted for each class, when precision is higher that means the false positive is fewer for example in Tamazight class. The model is exact.

In Recall matrix that calculates the predicted correctly among all actual classes for each class, when precision is higher that means the false negative is fewer for example in sport class. and all classes the recall score is high.

Enhancement ways to get better results:

- Using neural network model to classify the data with more layers.
- Try different ways in feature extraction by using embedding layer of NN model.
- Using Ensemble models such as Random Forest