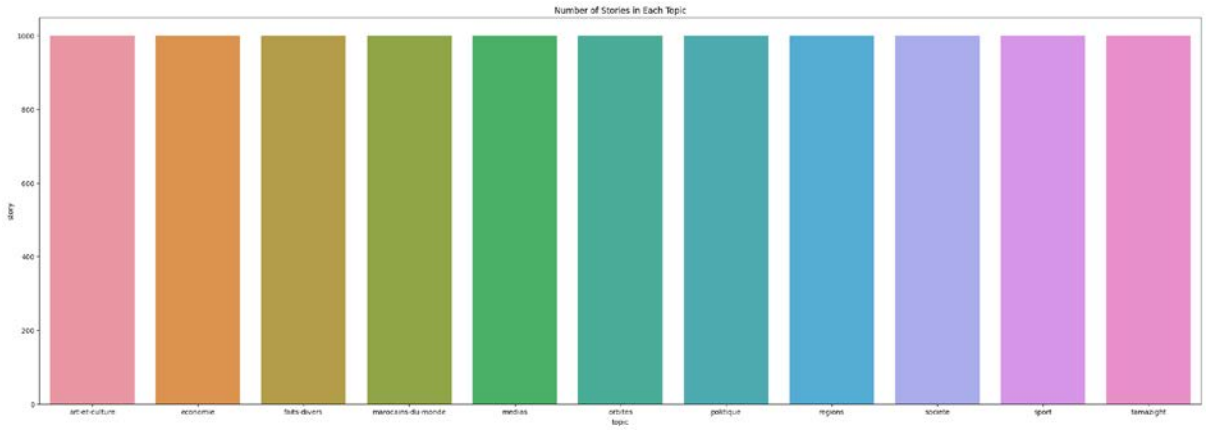
WideBot Task 2: EDA

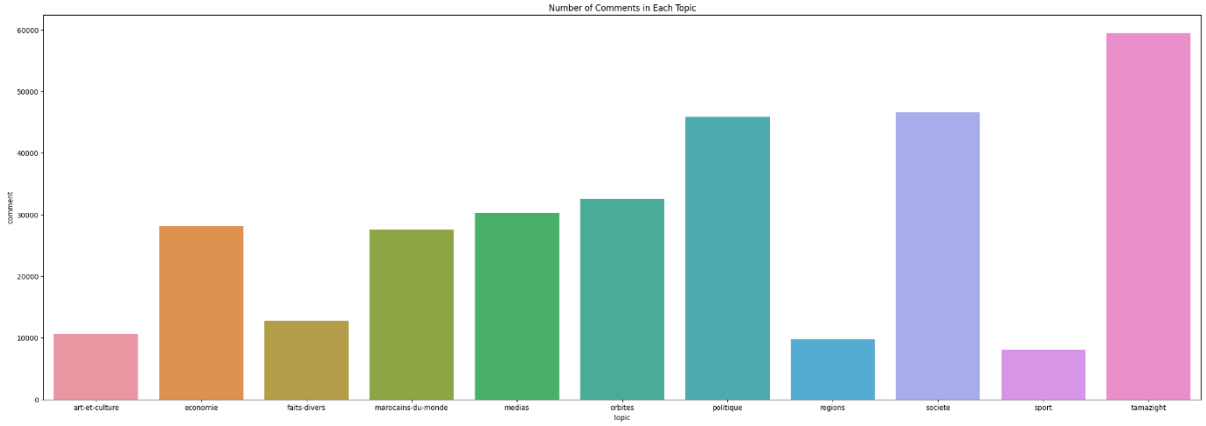
By: Aisha Hagar

Number of Examples per Class

For the stories dataset, there is an equal number of stories in each topic.



The comments dataset is imbalanced. From the below figure, we can see that most comments were on Tamazight stories and the least comments were on the sport stories.



Most common words

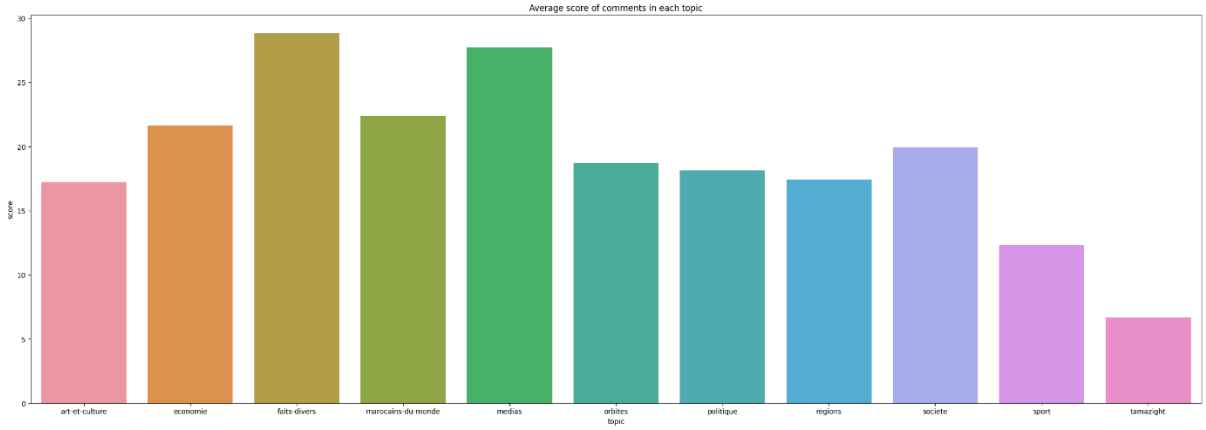
The table below shows the most common words in the stories dataset

	Word	Frequency
0	``	158355
1	خلال	10860
2	المغرب	8810
3	المغربية	8327
4	أنه	6672
5	كورونا	6032
6	الأمازيغية	5466
7	المغربي	5230
8	محمد	5133
9	حالة	5048

The table below shows the most common words in the comments dataset

	Word	Frequency
0	ان	102709
1	الله	85707
2	...	70887
3	المغرب	69766
4	``	53452
5	..	52879
6	الى	48942
7	او	41619
8	المغاربة	31855
9	يجب	28619

Average Score of comments in each topic



The above figure shows that Tamazight has the lowest average score despite having the highest number of comments.