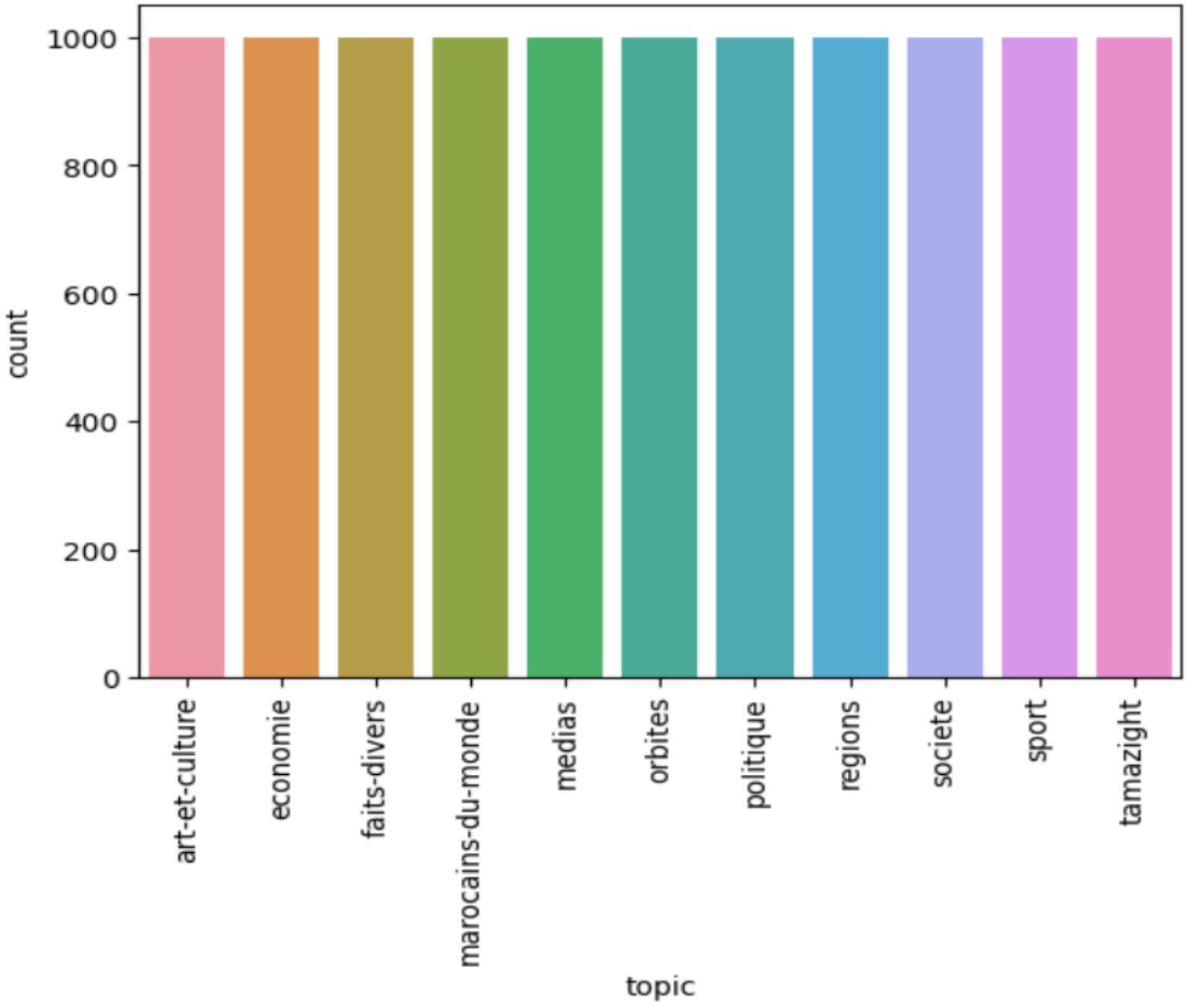
Exploratory Data Analysis (EDA) Report

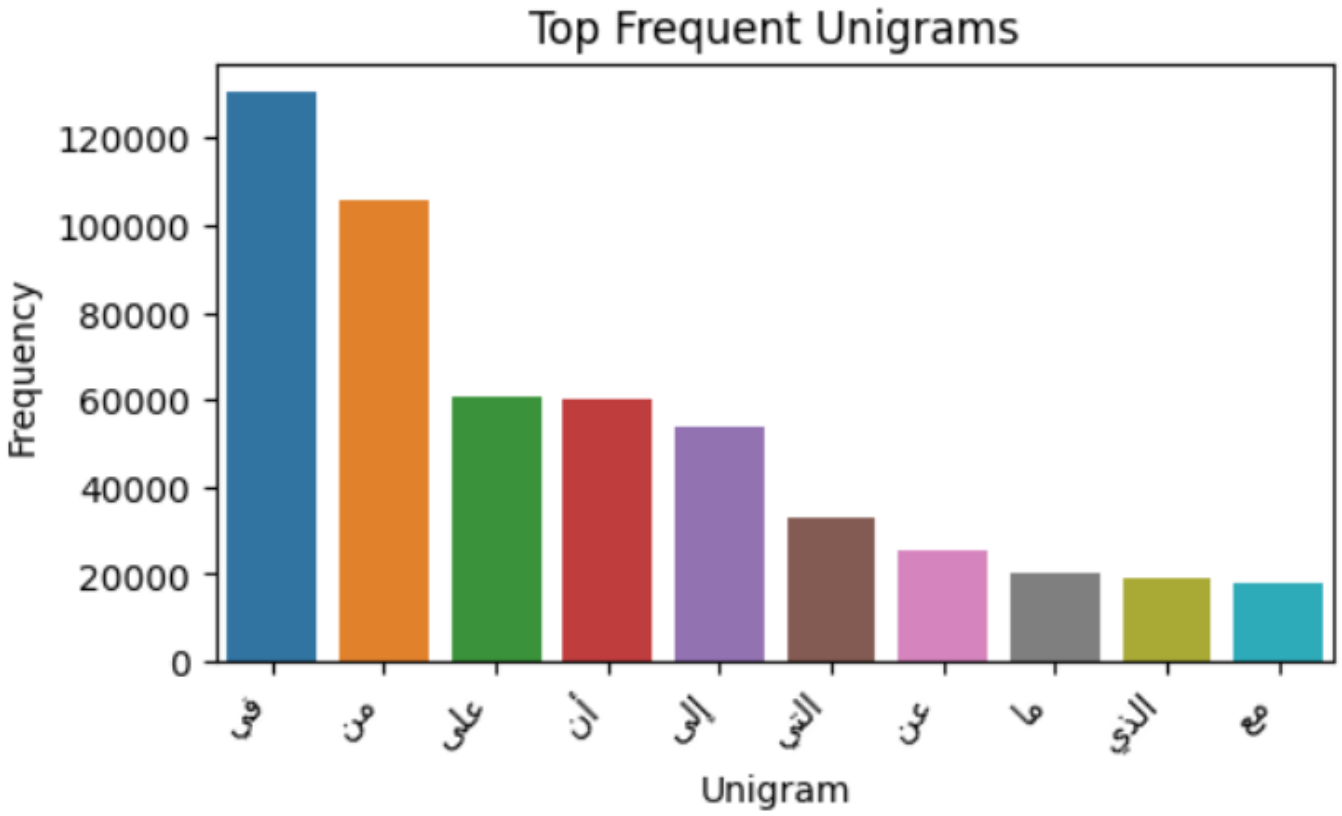
Insight 1: Number of Examples per Class.

- **The dataset is well-balanced, with each class having exactly 1000 examples. The following is the count of Stories in each class:**



Insight 2: Top Frequent Unigrams.

- **The most frequent unigrams (single words) in the dataset, along with their respective frequencies, are as follows:**



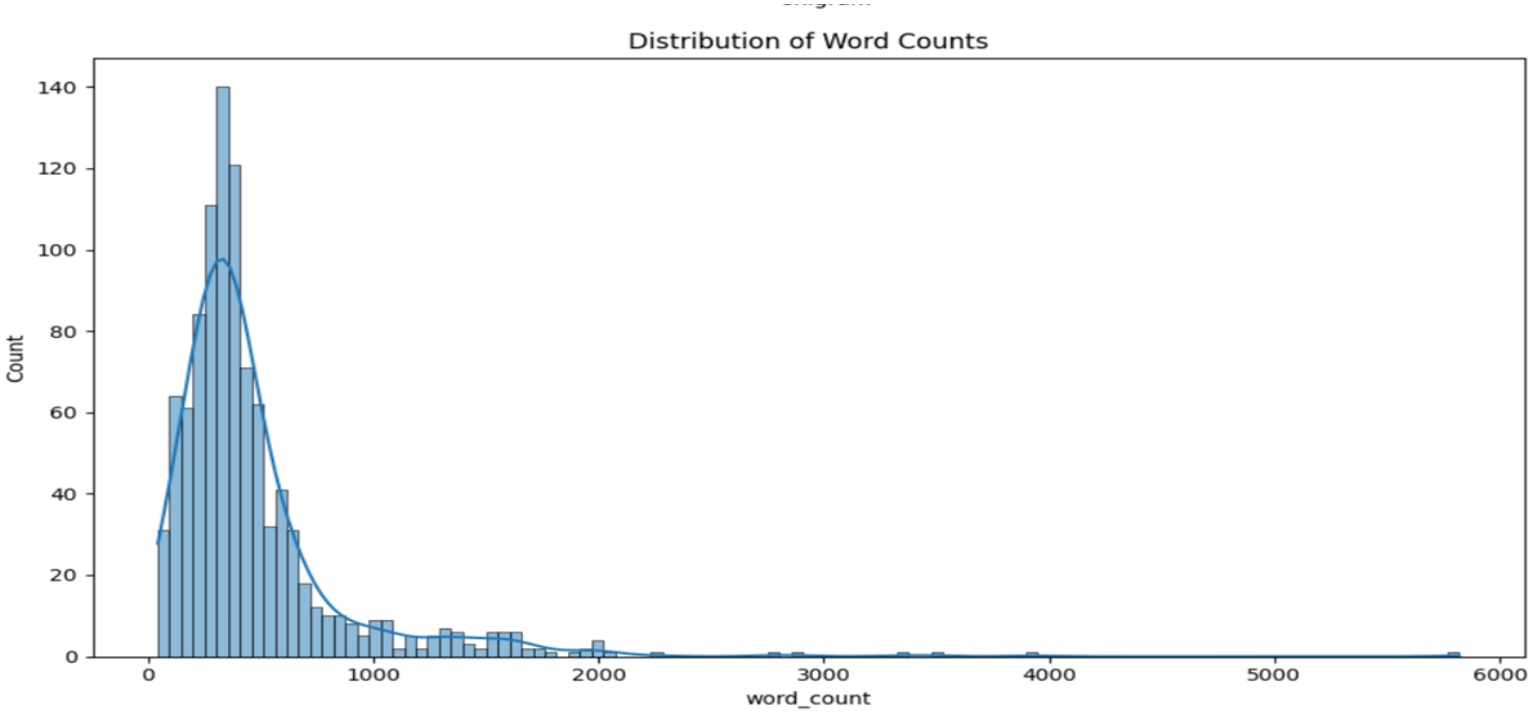
Insight 2: Top frequent unigrams

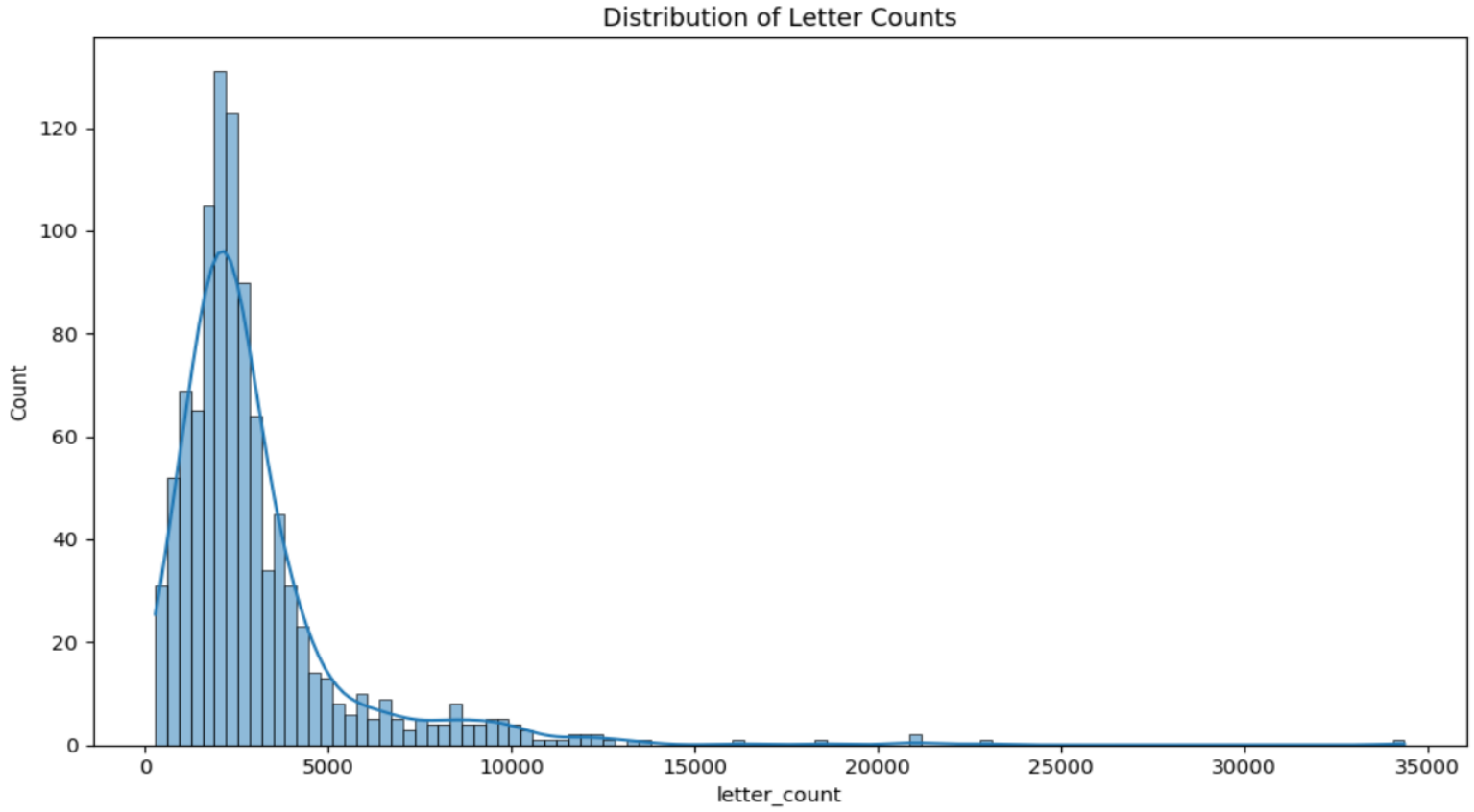
	Unigram	Frequency
0	130582	في
1	105589	من
2	60808	على
3	60166	أن
4	53806	إلى
5	33066	التي
6	25230	عن
7	20143	ما
8	19082	الذي
9	17956	مع

The unigram "في" appears most frequently, followed by "من" and "على." These insights indicate that certain words are highly common in the dataset and may not carry substantial discriminatory information.

Insight 3: Lengths of Examples **(word and Letter)**

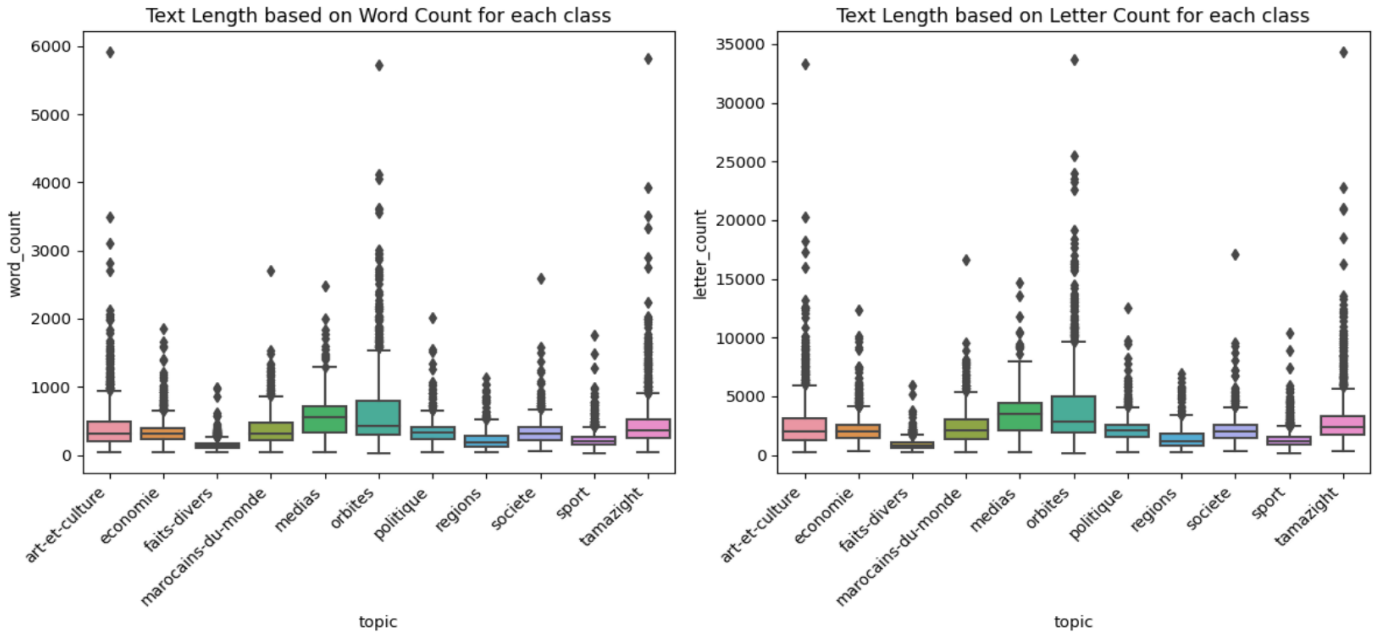
The distribution of the word count and letter count for the examples is as follows:





- **These insights reveal that the articles vary significantly in length, with the number of words ranging from (500) to 2000 and the number of letters ranging from (2500) up to 10000. Understanding the article lengths can help in choosing appropriate text representation techniques and modeling strategies.**

The next graphs show that and, we see that we found **outliers** in all classes:



Insight 4: Lengths of Examples **(word and Letter) for story only.**

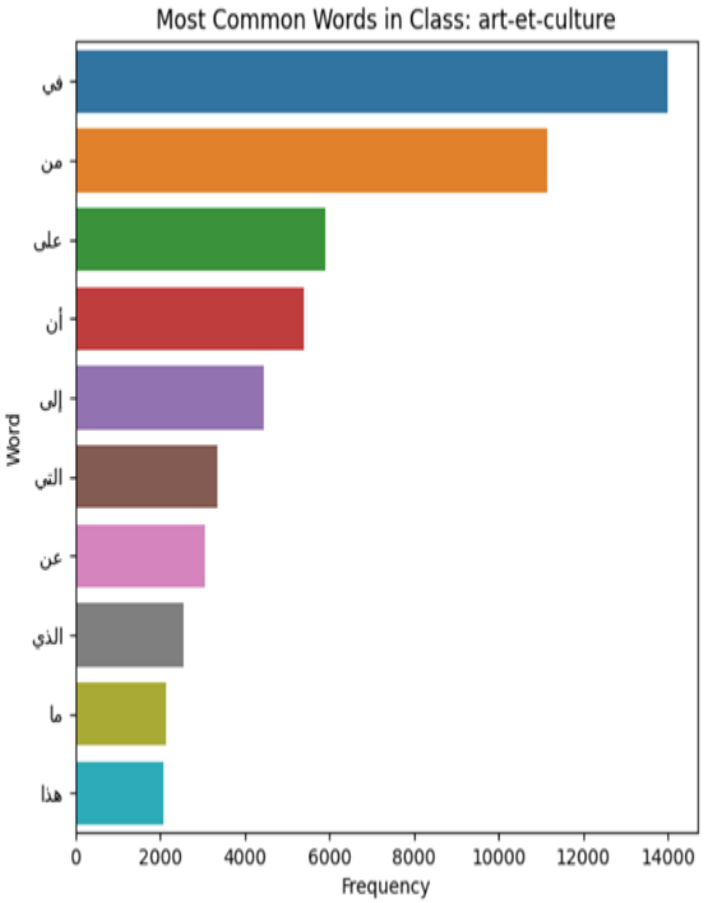
Word:

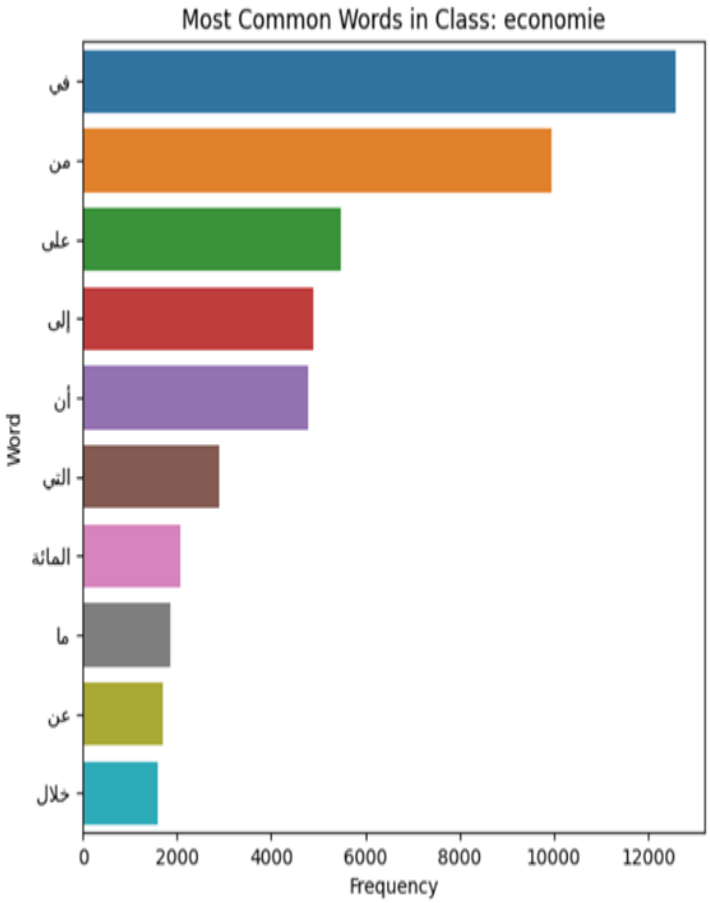
- Minimum Word Count: 200 - **Maximum Word Count:** 1255 - **Average Word Count: Approximately**:[556]

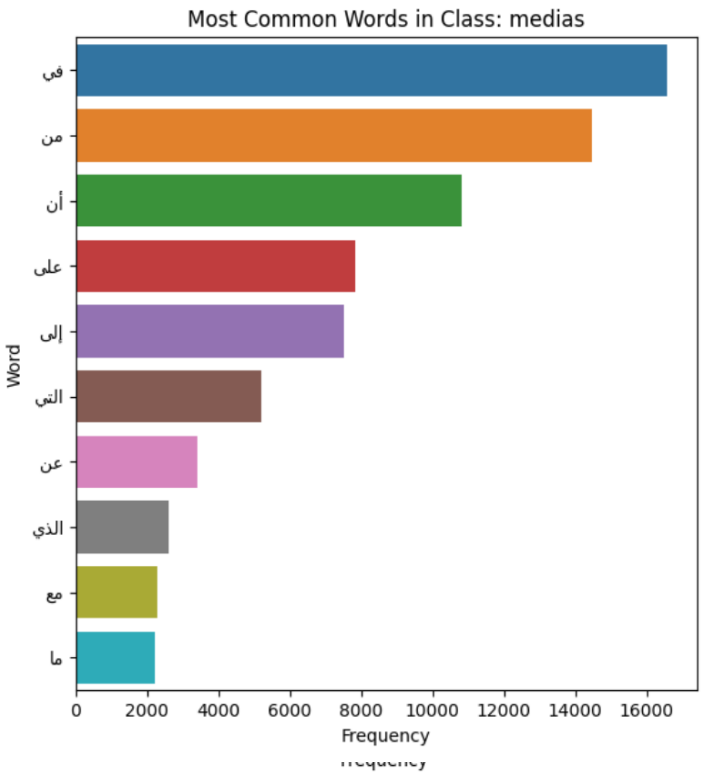
Letter: Minimum Letter Count: 1238 - **Maximum Letter Count:** 7773 - **Average Letter Count: Approximately** [3287]

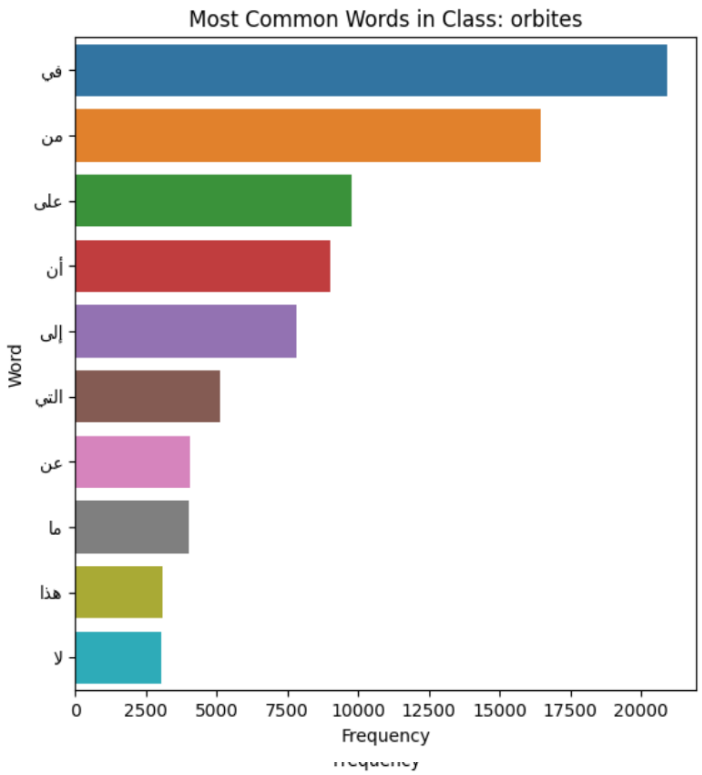
- These insights reveal that the articles vary significantly in length, with the number of words ranging from 200 to 1255 and the number of letters ranging from 1238 to 7773.

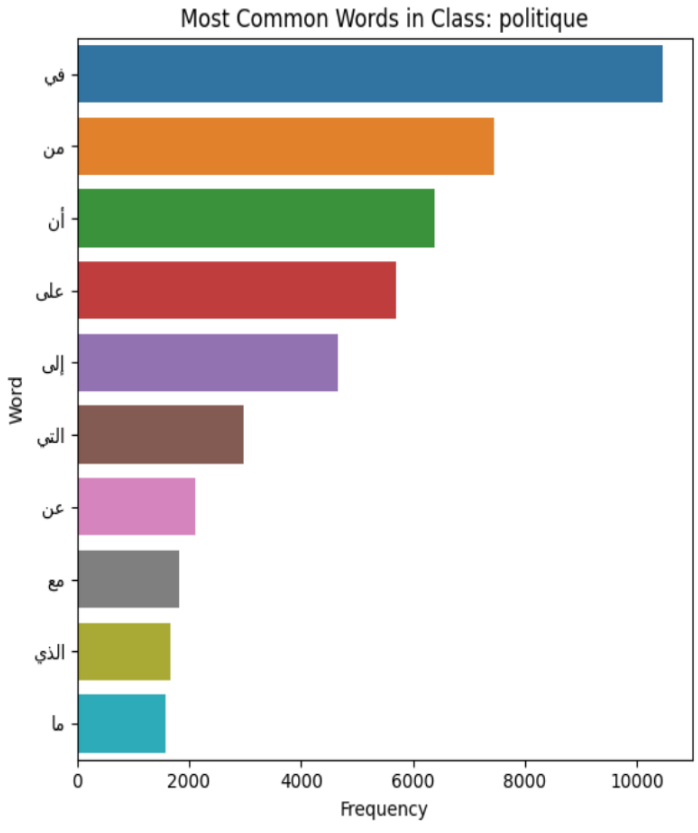
Insight 5: Most Common Words in each class

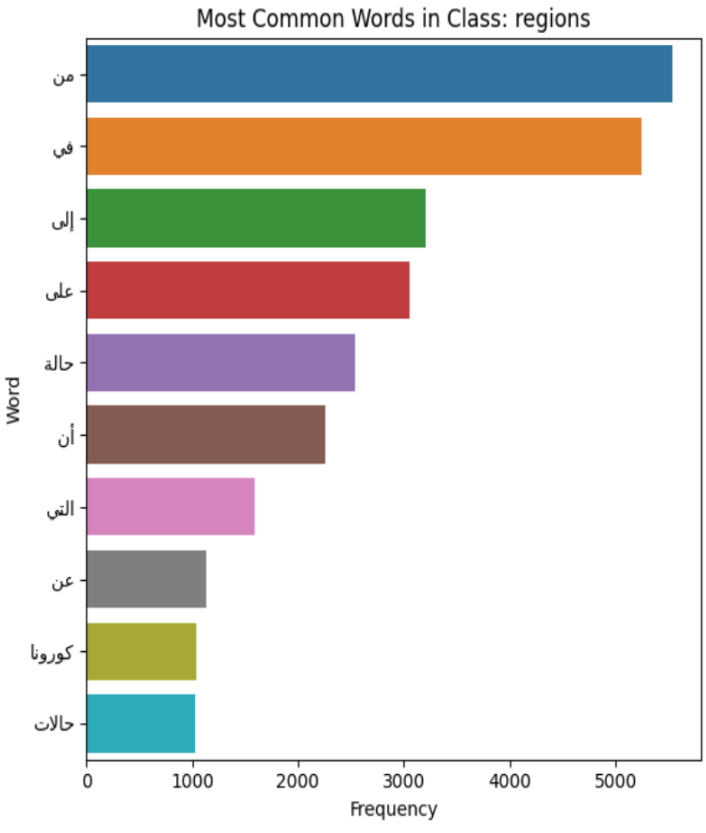


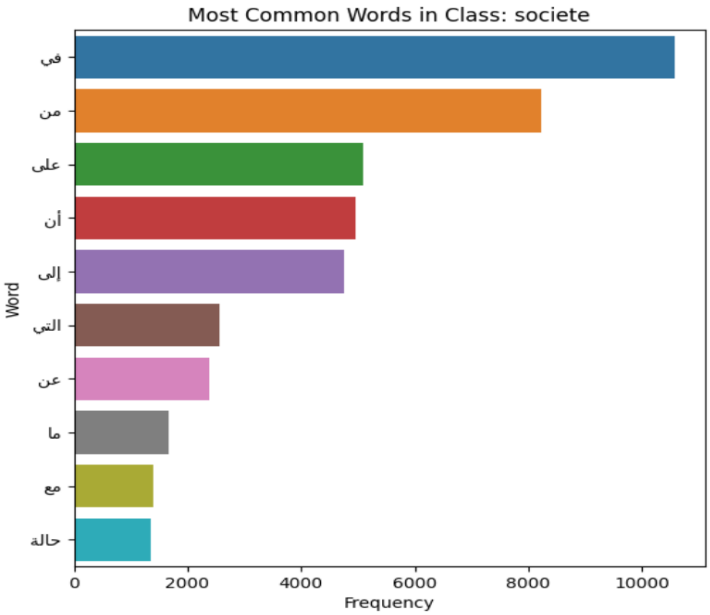


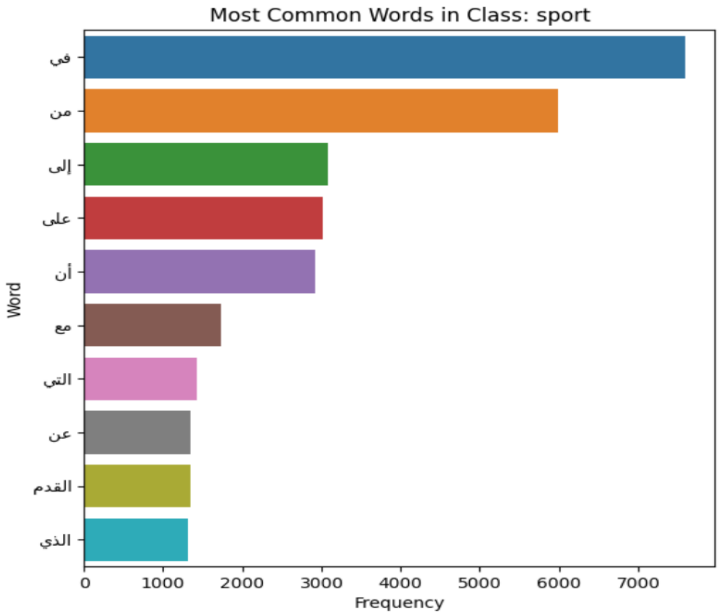

Most Common Words in Class: medias
في
من
أن
على
إلى
التي
عن
الذي
مع
ما
Word
0
2000
4000
6000
8000
10000
12000
14000
16000
Frequency

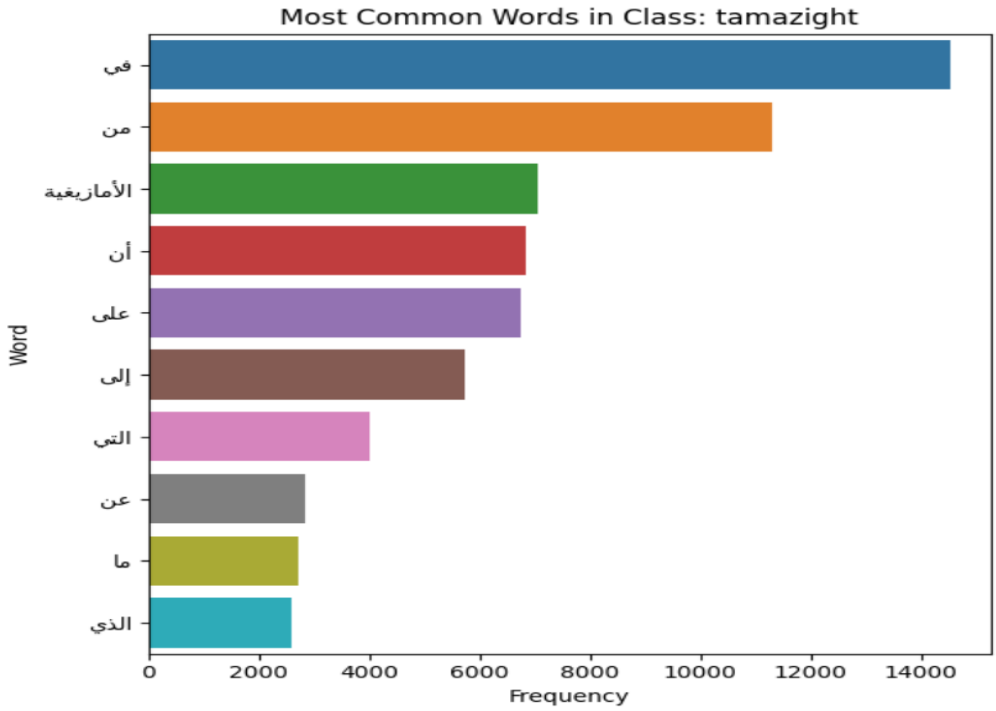

Most Common Words in Class: orbites
في
من
على
أن
إلى
التي
عن
ما
هذا
لا
Word
0
2500
5000
7500
10000
12500
15000
17500
20000
Frequency


Most Common Words in Class: politique
في
من
أن
على
إلى
التي
عن
مع
الذي
ما
Word
0
2000
4000
6000
8000
10000
Frequency


Most Common Words in Class: regions
من
في
إلى
على
حالة
أن
التي
عن
كورونا
حالات
Word
0
1000
2000
3000
4000
5000
Frequency







•The chart provides valuable insights into the most frequent words associated with each class, allowing us to identify the distinctive features of each category.

•Classes with specific domains, such as "art-et-culture" or "sports," tend to have domain-specific vocabulary in their most common words.

•Some words may appear in multiple classes, indicating potential overlapping themes or topics.