Exploratory Data Analysis

HESPRESS KAGGLE DATASET

Date handed in: 25 / 07 / 2023.

PROPLEM FORMULATION

Text classification and sentiment analysis is a crucial natural language processing (NLP) technique that involves analyzing Arabic text to determine its emotional tone, classifying it as positive, negative, or neutral. In this project, we aim to leverage this technique to develop classification for Arabic text. This is used in organizations for a variety of applications, including:

- Understanding Public Opinion: Analyzing sentiments from comments can help in understanding public opinion on various topics like education, crime, politics, etc.
- Brand Perception: Organizations can gauge public sentiment towards their brands or products by analyzing comments and stories related to them.
- Election Result Prediction: The sentiment analysis results can be aggregated to predict the sentiment of Moroccans towards certain topics and potentially predict election outcomes.

Overall, it's a great way for businesses to connect and understand what their customers want.

Dataset

Introduction:

The dataset under investigation aims to contribute to the development of Arabic Natural Language Processing (NLP) by providing valuable insights into public sentiment and opinions. The dataset comprises over 10,000 stories scraped from Hespress, a news website, and includes crucial details such as author information, publishing dates, and topics. Additionally, it includes more than 300,000 comments associated with the stories, each accompanied by a reader's score.

Dataset Overview:

- Total Stories: 10,000: -

	id	title	date	author	story	topic
0	9de52a46055311eb8949646e69d991ea	وزارة التربية: لا تغيير في الغلاف الزمني للأمازيغية	السبت 03 أكتوبر 00:02 - 2020	هسبريس من الرباط	قالت وزارة التربية الوطنية والتكوين المهني والتعليم العالي والبحث العلمي، قطاع التربية الوطنية، ...	tamazight
1	9ee74b02055311ebb757646e69d991ea	تغييب" تدريس الأمازيغية يغضب" نقابات بتزنيت	الاثنين 28 شتنبر 09:13 - 2020	رشيد بيجيكن من أكادير	طالبت الكتابة الإقليمية للجامعة الوطنية للتعليم -التوجه الديمقراطي- بتزنيت المدير الإقليمي لوزار...	tamazight
2	9fed7812055311eb9158646e69d991ea	مرصد يستنكر رفض قبول أستاذة أمازيغية بمدرسة	الاثنين 28 شتنبر 00:41 - 2020	هسبريس من الرباط	أفاد المرصد الأمازيغي للحقوق والحريات أنه توصل بمعطيات تفيد بأن أستاذة للغة الأمازيغية حديثة الت...	tamazight
3	a0e39038055311eb8f78646e69d991ea	نص امتحان موحد يثير حنق فعاليات أمازيغية	السبت 26 شتنبر 17:28 - 2020	هسبريس من الرباط	سجلت فعاليات مدنية أمازيغية عديدة امتعاضها من نص امتحان في اللغة العربية لنيل شهادة التعليم الاب...	tamazight
4	a1d90814055311eb879e646e69d991ea	عادل تيزنيت" يتحف المغاربة بالموسيقى الأمازيغية	السبت 26 شتنبر 15:50 - 2020	هسبريس من الرباط	من داخل سيارة بسيطة يركنها بالمدخل الشمالي لمدينة تيزنيت، تمكن الفنان المغربي "عادل تيزنيت" من ا...	tamazight

Figure 1 (Sample of story's data)

- Total Comments: 300,000

	postId	comment	score	topic
0	9ee74b02055311ebb757646e69d991ea	بعيدا عن بعض التعليق العنصرية ،الامازيغية لغة و ثراث عريق و ملك لجميع المغاربة يجب الرقي بها لان...	-36	tamazight
1	9ee74b02055311ebb757646e69d991ea	... كان اكبر خطء ارتكبه المغرب نتيجة الخريف العربي هو الاعتراف بلهجة مغربية كلغة وهو ما يؤسسس لصراع	66	tamazight
2	9ee74b02055311ebb757646e69d991ea	كا نتوقع من شي وحدين بلا حشما بلا حشوم غادي قولو آش عطاتنا الأمازيغية، واش ماشي من حق الناس دافع...	-47	tamazight
3	9ee74b02055311ebb757646e69d991ea	ارجوا من الاخوة المسؤولين احترام اللغة الأمازيغية .لانها دعامة كبيرة لهويتنت وشخصيتها وتميزنا	-32	tamazight
4	9ee74b02055311ebb757646e69d991ea	... تدريس الامازيغية مضيعة للوقت والمال وفتح باب الفتنة بين المغاربة والتاسيس لصراع لا طاءل منه بين	55	tamazight

Figure 2 (Sample of comment's data)

- Attributes: Story ID, Author, Publishing Date, Topic, Comment ID, Comment Text, Reader's Score

Key Insights:

1. Class Distribution
2. Text Analysis
3. Story Length Analysis
4. Public Sentiment for Specific Topics

TECHNIQUES & DISCUSSION

We started working on the stories' dataset.

Non-Null values: -

The data has no null values as shown in figure 3.

```
Int64Index: 11000 entries, 0 to 999
Data columns (total 6 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   id      11000 non-null  object
 1   title   11000 non-null  object
 2   date    11000 non-null  object
 3   author  11000 non-null  object
 4   story   11000 non-null  object
 5   topic   11000 non-null  object
dtypes: object(6)
```

Figure 3

Number of duplicate entries: -

The data has no duplicated entries.

Number of stories per topic: -

As shown in figure 4, we have almost 1000 stories per topic, our dataset is perfectly balanced.

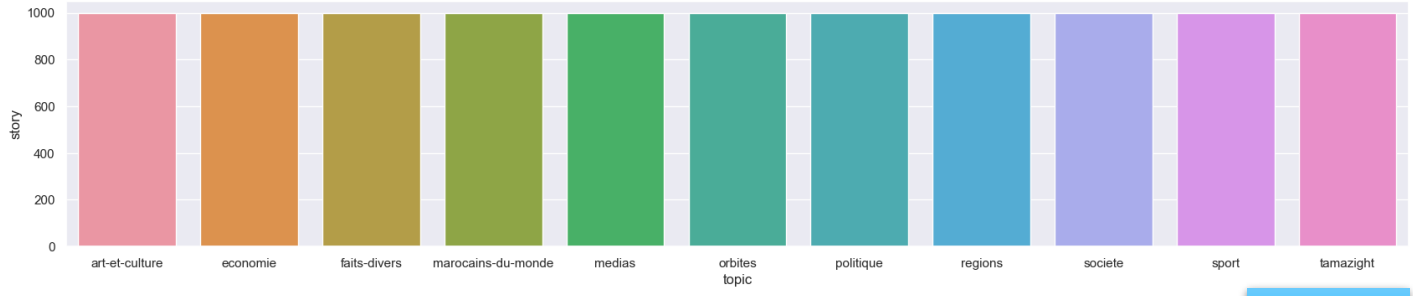

Figure 4

Lengths of examples in words and letters: -

For each entry, we calculated the words and letters count and added it to the data frame in the last 2 columns, as shown in figure 5.

	id	title	date	author	story	topic	letters_count	words_count
0	9de52a46055311eb8949646e69d991ea	وزارة التربية: لا تغيير في الغلاف الزمني للأما...	السبت 03 أكتوبر 00:02 - 2020	هسبريس من الرباط	قالت وزارة التربية الوطنية والتكوين المهني وال...	tamazight	472	83
1	9ee74b02055311ebb757646e69d991ea	تغييب" تدريس الأمازيغية يغضب نقابات بتزنيت"	الاثنين 28 شتنبر 09:13 - 2020	رشيد بيجيكن من أكادير	طالبت الكتابة الإقليمية للجامعة الوطنية للتعلي...	tamazight	1402	246
2	9fed7812055311eb9158646e69d991ea	مرصد يستنكر رفض قبول أستاذة أمازيغية بمدرسة	الاثنين 28 شتنبر 00:41 - 2020	هسبريس من الرباط	أفاد المرصد الأمازيغي للحقوق والحريات أنه توصل...	tamazight	1143	218
3	a0e39038055311eb8f78646e69d991ea	نص امتحان موحد يثير حنق فعاليات أمازيغية	السبت 26 شتنبر 17:28 - 2020	هسبريس من الرّباط	سجلت فعاليات مدنية أمازيغية عديدة امتعاضها من ...	tamazight	514	96
4	a1d90814055311eb879e646e69d991ea	عادل تيزنيت" يتحف المغاربة بالموسيقى الأمازيغية	السبت 26 شتنبر 15:50 - 2020	هسبريس من الرباط	من داخل سيارة بسيطة يركنها بالمدخل الشمالي لمد...	tamazight	1437	274

Figure 5

Figure 6 illustrates both the average words and letters count per class in the stories' dataset.

Figures 7 and 8 illustrate the box plot of letters and words per class respectively. Note that we removed the outliers to make it more visible and understandable.

topic	average_letters_count	average_words_count
art-et-culture	2139.847	426.099
economie	1774.603	344.926
faits-divers	757.556	149.231
marocains-du-monde	1931.584	374.559
medias	2778.525	545.078
orbites	3238.874	642.337
politique	1753.405	336.144
regions	1160.005	223.210
societe	1705.422	331.992
sport	1135.691	227.697
tamazight	2431.089	471.237

Figure 6

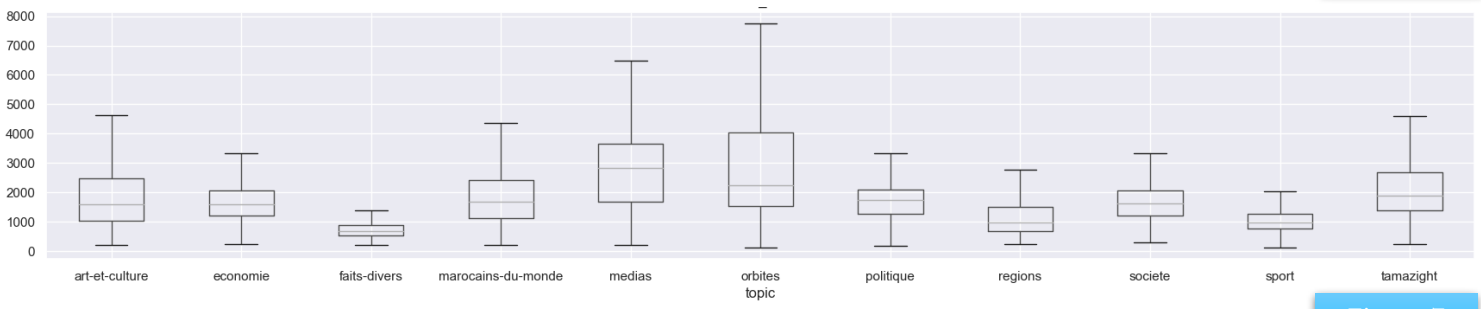

Figure 7

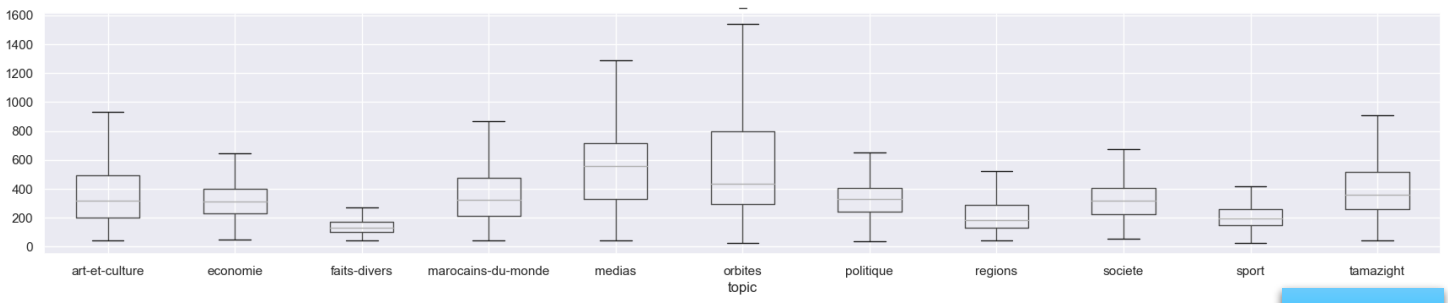

Figure 8

Top frequent n-grams generally: -

Figure 9 illustrates the top 10 frequent bigram generally in the dataset.

	Bigram	Frequency
0	من أجل	5733
1	إلى أن	5654
2	من خلال	3604
3	في المائة	3103
4	وهو ما	2904
5	من طرف	2704
6	في هذا	2690
7	فيروس كورونا	2546
8	على أن	2319
9	في إطار	2216

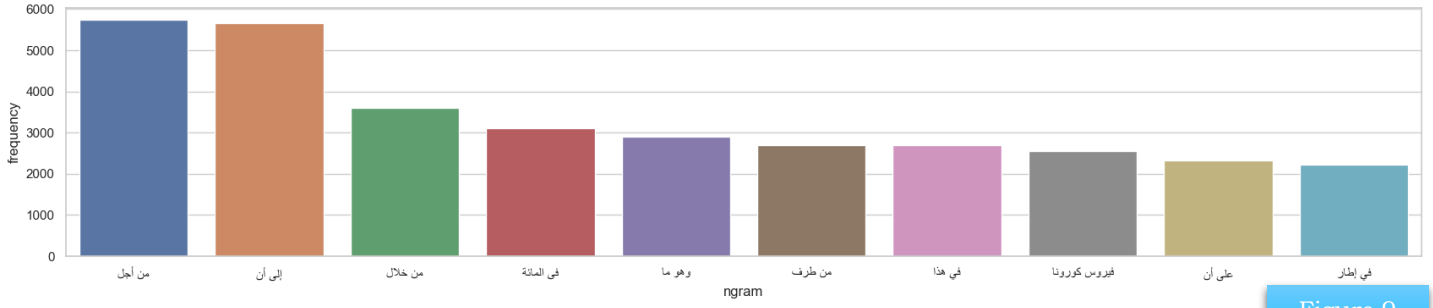

Figure 9

Figure 10 illustrates the top 10 frequent trigram generally in the dataset.

	Trigram	Frequency
0	مشيرا إلى أن	1131
1	الملك محمد السادس	1105
2	في المائة من	854
3	لجريدة هسبريس الإلكترونية	808
4	فيروس كورونا المستجد	796
5	في تصريح لهسبريس	682
6	تصريح لجريدة هسبريس	675
7	في تصريح لجريدة	659
8	النيابة العامة المختصة	585
9	بفيروس كورونا المستجد	580

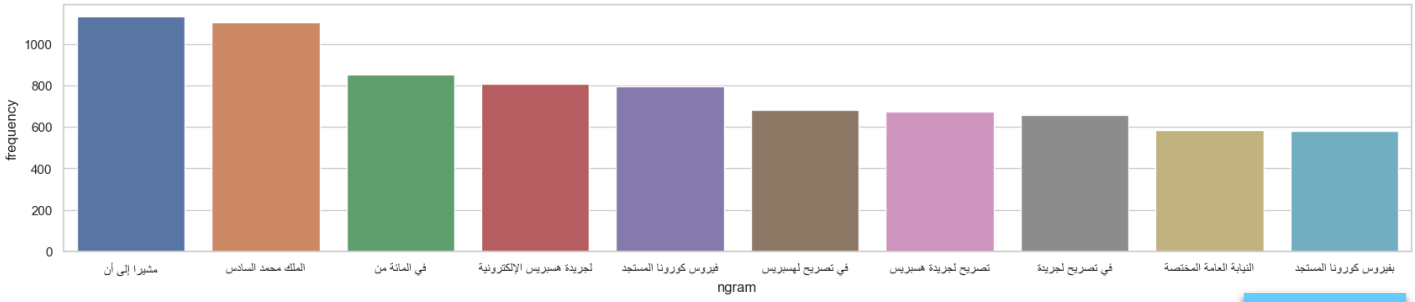

Figure 10

In the following page, figures 11 and 12 illustrates the most bigram and trigram per class.

Figure 11

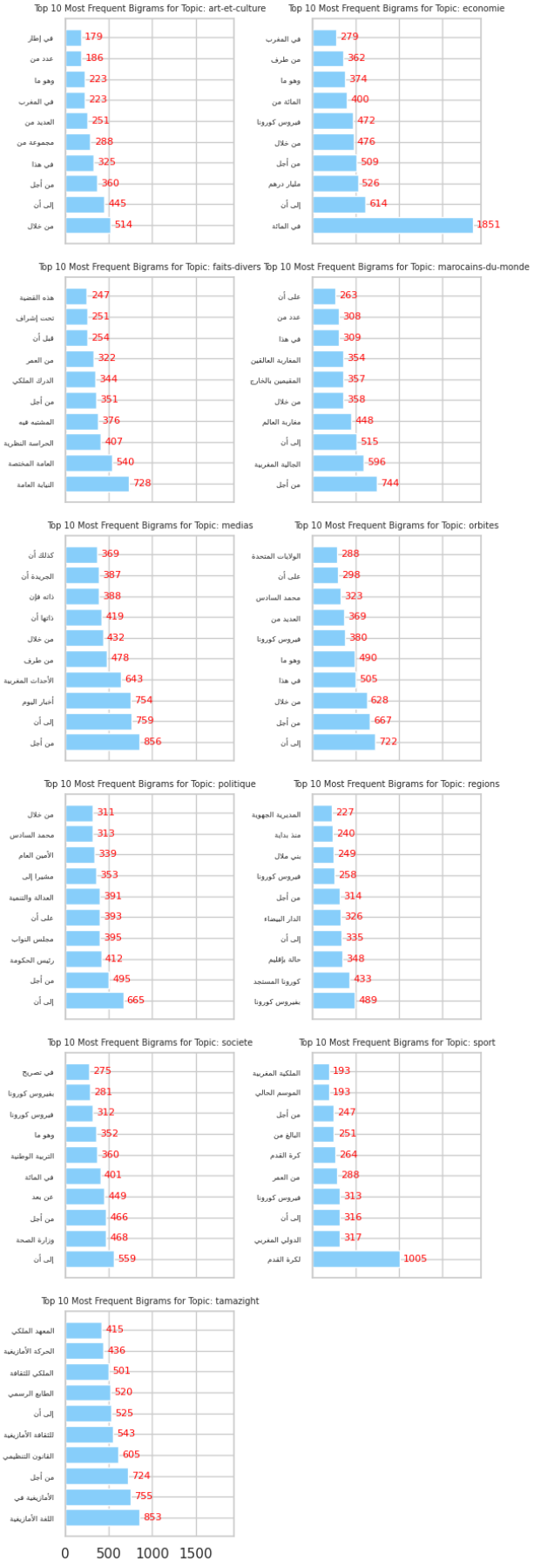

Figure 12