Arabic Tweets Sentiment Analysis

Amjid Khurwat - Johnny Lamaa - Nacer Boudad

Data

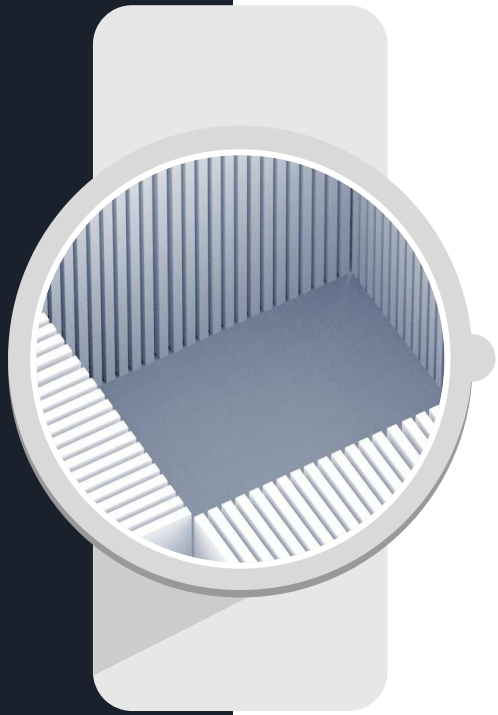
In total, we have acquired a dataset consisting of 11211 data points/tweets. Each tweet is in arabic and classified into one of 3 classes: 0 for normal - 1 for hateful - 2 for abusive -3 for bullying

We assembled the final dataset from 4 different public repositories:

1. Arabic-Hate-Speech by Raghdash.
2. MLMA-Hate-Speech by HKUST-KnowComp
3. Arabic Levantine Hate Speech Detection by Haithem Hermessi
4. Terrorism and Jihadism Speech Detection by Haithem Hermessi

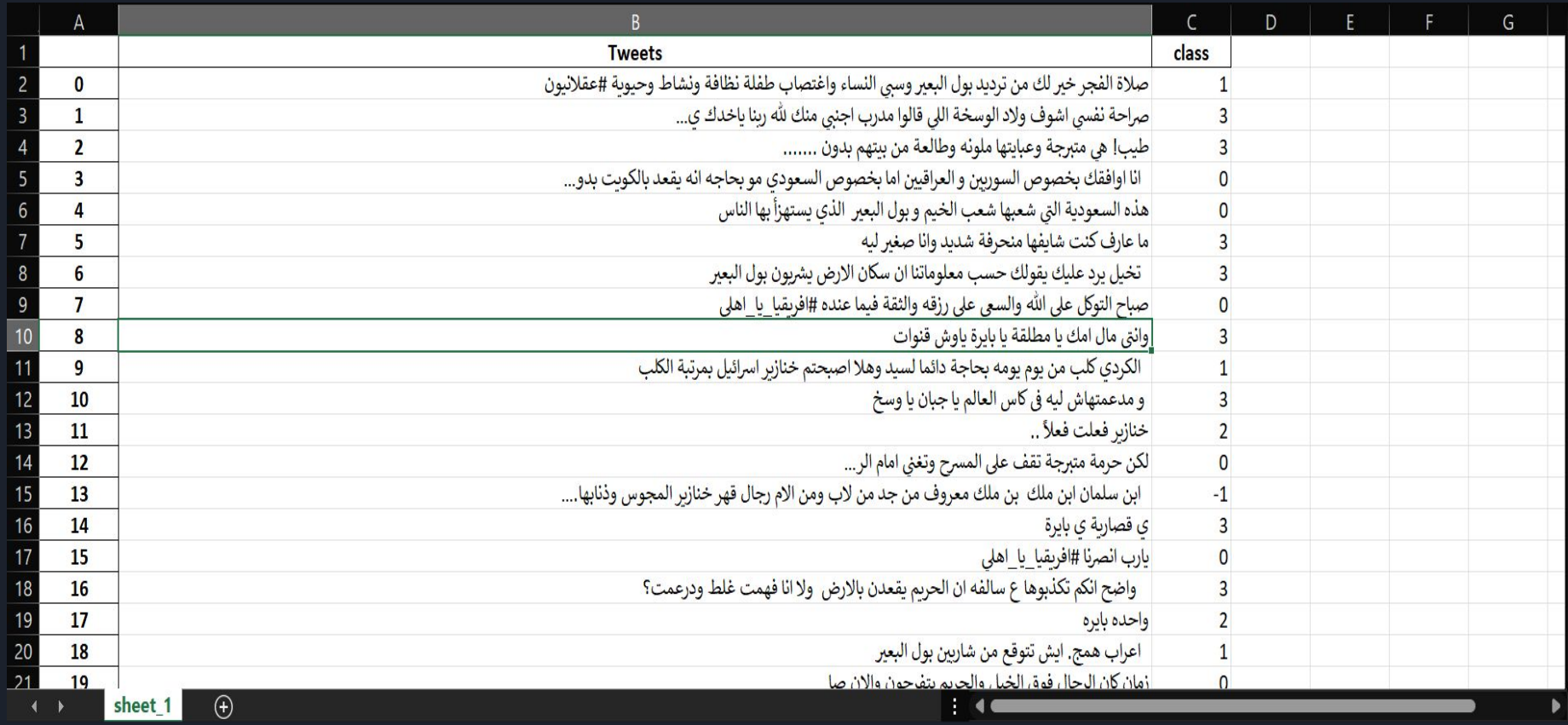
Data Preprocessing

Some of the datasets had french written tweets, deleted or invalid tweets (through the twitter API), so we had to select only the arabic tweets from all the labeled tweets to get 11211 well classified tweets.



We also found an arabic text preprocessing (DSAraby) library by Saobou to help with the preprocessing of the tweets through:

1. Transliteration
2. Text Normalization
3. Arabic Stop Words
4. Dealing with Hashtags
5. Dealin with Emojis
6. Removing Links
7. RT, CC, Mentions
8. Remove Digits, dash, punctuation marks and any other marks



	A	B	C	D	E	F	G
1		Tweets	class				
2	0	صلاة الفجر خير لك من ترديد بول البعير وسبي النساء واغتصاب طفلة ونشاط وحيوية #عقلانيون	1				
3	1	صراحة نفسي اشوف ولاد الوسخة اللي قالوا مدرب اجنبي منك لله ربنا ياخدك ي...	3				
4	2	طيب! هي متبرجة وعبايتها ملونه وطالعة من بيتهم بدون	3				
5	3	انا اوافقك بخصوص السوريين و العراقيين اما بخصوص السعودي مو بحاجه انه يقعد بالكويت بدو...	0				
6	4	هذه السعودية التي شعبها شعب الخيم و بول البعير الذي يستهزأ بها الناس	0				
7	5	ما عارف كنت شايفها منحرفة شديد وانا صغير ليه	3				
8	6	تخيل يرد عليك يقولك حسب معلوماتنا ان سكان الارض يشربون بول البعير	3				
9	7	صباح التوكل على الله والسعي على رزقه والثقة فيما عنده #افريقيا_يا_اهلى	0				
10	8	وانتى مال امك يا مطلقة يا بايرة ياوش قنوات	3				
11	9	الكردي كلب من يوم يومه بحاجة دائما لسيد وهلا اصبحتم خنازير اسرائيل بمرتبة الكلب	1				
12	10	و مدعمتهاش ليه فى كاس العالم يا جبان يا وسخ	3				
13	11	خنازير فعلت فعلاً ..	2				
14	12	لكن حرمة متبرجة تقف على المسرح وتغني امام الر...	0				
15	13	ابن سلمان ابن ملك بن ملك معروف من جد من لاب ومن الام رجال قهر خنازير المجوس وذنابها....	-1				
16	14	ي قصارية ي بايرة	3				
17	15	يارب انصرنا #افريقيا_يا_اهلي	0				
18	16	واضح انكم تكذبوها ع سالفه ان الحريم يقعدن بالارض ولا انا فهمت غلط ودرعمت؟	3				
19	17	واحده بايره	2				
20	18	اعراب همج. ايش تتوقع من شارين بول البعير	1				
21	19	زمان كان الرجال فوق الخيل والحريم يتفرجون والان صا	0				

sheet_1

Data snippet

Models

WE will experiment with three main machine learning models that would be the most efficient for our case:

1. SVM
    a. Polynomial Kernel
    b. RBF Kernel
2. Logistic Regression Model

3. We will also attempt to build a small neural network although our data size is small (11211 points)