Sentiment Analysis task

Introduction :

The task is to make a pipeline that classifies the tweets to pos/neg/neu and to do a 10 cv when reporting accuracy and f1 score

* As mentioned in the task our first approach will be by performing **Tf-IDF** and using various **classifiers** to get the results
* Second, we will train a **Bert** model on our data and report the results
* Finally, we will use an endpoint that is well trained on Arabic tweets, and report the results.

The data:

* The data provided for the task 2700 rows of arabic tweets labeled pos/neg/neu

We split this data to 80/20 for training and testing

* I used more data to enrich our data for getting better results
  + ArSaS <http://lrec-conf.org/workshops/lrec2018/W30/pdf/22_W30.pdf>
    - This data is nearly 20k rows and by dropping the Mixed label we get around 18k rows
  + ASTD <https://github.com/mahmoudnabil/ASTD.git>
    - This dataset contains an Objective label and by dropping it we get NEG 1642 / NEUTRAL 805 / POS 777 around 2200l row

The tools:

1. SK-learn
2. Pandas
3. Numpy
4. Huggingface
5. Regex
6. Pyarabic used for manipulating Arabic letters
7. Emoji used for processing emoji
8. Pystemmer used for performing stemming <https://github.com/snowballstem/pystemmer>

The process:

Machine learning Notebook

* First we have conduct our baseline experiment by processing the data throw tfidf and getting our baseline results
  + The baseline results was **accuracy: 58% —-- f1score: 53%**

The models test on are :

* + - LinearSVC
    - MultinomialNB
    - BernoulliNB
    - SGDClassifier
    - DecisionTreeClassifier
    - RandomForestClassifier
    - KNeighborsClassifier

And our best model was LinearSVC

* Then we changes the strategy of the Tf-Idf to Char and used up to 6 ngram

We got results **accuracy: 62% —-- f1score: 61%**

* We did some preprocessing to the dataset but we got no improvement

The cleaning included removing English words, hashtags, repeated words, symbols, mentions and replacing some rarely used Arabic character

* Then we enriched our data with more data and applied the preprocessing on them and we get the best results from this notebook to

**accuracy: 63% —-- f1score: 62%**

AS the difference between the dev and the Test is very large that suggests that they have a different distribution so we will not use extra data in the DL approach.

Bert Notebook

* We did the cleaning to our data (same cleaning in the pervious approach)
* We used MARBERT endpoint on with the following parameter:
  + training\_args.lr\_scheduler\_type = 'cosine'
  + training\_args.evaluate\_during\_training = True
  + training\_args.adam\_epsilon =1e-8
  + training\_args.learning\_rate = 5e-05
  + training\_args.fp16 = True
  + training\_args.per\_device\_train\_batch\_size = 16 #64
  + training\_args.per\_device\_eval\_batch\_size = 16 # 64
  + training\_args.gradient\_accumulation\_steps = 2
  + training\_args.num\_train\_epochs= 10
  + training\_args.warmup\_steps = 500
  + training\_args.evaluation\_strategy = EvaluationStrategy.EPOCH
  + training\_args.logging\_steps = 200
  + training\_args.save\_steps = 100000
  + training\_args.seed = 50
  + training\_args.disable\_tqdm = False
* We were able to get **73%** as a result of our training

Camel Tool notebook:

* Our last approach was to use an endpoint for inference so we used the Camel tool for tweets

<https://huggingface.co/CAMeL-Lab/bert-base-arabic-camelbert-da-sentiment?text=%D9%83%D9%84%D8%A7%D9%85+%D8%A7%D9%84%D9%86%D8%A7%D8%B3+%D9%84%D8%A7+%D8%A8%D9%8A%D9%82%D8%AF%D9%85+%D9%88%D9%84%D8%A7+%D9%8A%D8%A3%D8%AE%D8%B1+%D9%83%D9%84%D8%A7%D9%85+%D8%A7%D9%84%D9%86%D8%A7%D8%B3+%23%D8%B9%D9%86+%23%D8%A7%D9%84%D9%86%D9%8A%D8%B2%D9%83+%D9%85%D8%B4+%D8%A3%D9%83%D8%AA%D8%B1>

* The results from this experiment were 71% accuracy and 70%

Conclusion:

For investigating the reasons why we got poor accuracy I checked the annotation of the dataset given.

I found some inconsistent annotation and other that I dont agree with I provided some examples

Tweets :

* ' -هتعمل اية بكرة ال Christmas ؟ -هتفرج علي Home alone ✌ ',neg
* ' الاهتمام و التقدير مفتاح اى حد ',pos
* ' مشوفتش لاعب نشط ع تويتر زي Ferdinand. ',pos
* ' @SeRnGaA وربنا دعيت بضمير ده انا كنت مصلي الفجر :( ',pos

Challenges:

The data I get from the internet had a different distribution than the provided data so I wasnt be able to enrich the data.

Regefernces :

https://www.kaggle.com/mksaad/sentiment-analysis-in-arabic-tweets-using-sklearn

<https://www.kaggle.com/asalhi/arabic-sentiment-analysis-2nd-place-winning-code/notebook>

<https://huggingface.co/UBC-NLP/MARBERT?text=%D8%A7%D9%84%D9%84%D8%BA%D8%A9+%D8%A7%D9%84%D8%B9%D8%B1%D8%A8%D9%8A%D8%A9+%D9%87%D9%8A+%D9%84%D8%BA%D8%A9+%5BMASK%5D>.

<https://huggingface.co/CAMeL-Lab/bert-base-arabic-camelbert-da-sentiment?text=%D9%83%D9%84%D8%A7%D9%85+%D8%A7%D9%84%D9%86%D8%A7%D8%B3+%D9%84%D8%A7+%D8%A8%D9%8A%D9%82%D8%AF%D9%85+%D9%88%D9%84%D8%A7+%D9%8A%D8%A3%D8%AE%D8%B1+%D9%83%D9%84%D8%A7%D9%85+%D8%A7%D9%84%D9%86%D8%A7%D8%B3+%23%D8%B9%D9%86+%23%D8%A7%D9%84%D9%86%D9%8A%D8%B2%D9%83+%D9%85%D8%B4+%D8%A3%D9%83%D8%AA%D8%B1>