Sentimental Analysis of Arabic tweets

Summary

In this project ,Sentimental analysis is done on Arabic tweets .An Arabic-Bert-base model from hugging face is used,it was pretrained on ~8.2 Billion words which sum up to ~95GB of text.
The model achieved an accuracy of 90.16% and AUC of 0.966 on test set.

Motivation

Sentimental analysis is a crucial step in a lot of applications.
Whether you try to increase Customer satisfaction ,have better product consumption or to obtain important information regarding public opinion,Sentimental analysis helps you to efficiently achieve your goal.
Since there is a lack of applications that uses Arabic data-set in the market and since Arabic is the main language of more than 3.6% of world population , we wanted to help in enriching the Arabic based applications and encourage developers to do the same.

Procedure

1- Dataset was found on kaggle at this link:
https://www.kaggle.com/mksaad/arabic-sentiment-twitter-corpus

2- Some preprocessing was done before using BERT including:
- Normalize Unicode encoding
- Removing urls,@,trailing white spaces
- Add `[CLS]` and `[SEP]` to the beginning and end of each sentence
- set a max_length through truncating or padding

3- A Bert classifier is used , along with Adam optimizer and a learning rate scheduler.
4- Cross entropy was used as the objective function
5- Two performance metrics were used for evaluation : Accuracy and AUC

Dataset

The data-set contains 58K Arabic tweets (47K training, 11K test) tweets annotated in positive and negative labels. The data-set is balanced and collected using positive and negative emojis lexicon.
Data format: Tab-separated values TSV
Data files structure :the data folder has 4 tsv files, where negative and positive sentiments are separated for both training and testing
Link: https://www.kaggle.com/mksaad/arabic-sentiment-twitter-corpus

Results & conclusion

The model achieved an accuracy of 90.32% and AUC of 0.96 on validation set.
It achieved an accuracy of 90.16% and AUC of 0.966 on test set.
These results were achieved using only 2 epochs, we suggest that increasing the number of epochs could improve the accuracy of the model.